(Dis)inception: Dream Identification in Fiction

Srivastava Priyam, Morrissey Jack, Pandita Akhilesh

Abstract

Dream narratives have been long employed by authors to create gateways into the minds of narrators and characters, as well as alter the surrounding narratives in which they inhabit. Conventional methods like close reading rely on the judgement of the reader to separate dream narratives from non-dream narratives within texts, and this task becomes increasingly involved with large corpus of data. In this project, we present a framework to identify dream reports embedded within fictional narratives via the application of various machine learning models in combination with prevalent dream and literary theories. We explore these methods on three datasets consisting of clinical dream reports, English fiction, and dream reports within fiction. Our results are encouraging as we are able to demarcate dream narratives from other narratives through significant textual markers.

I. Introduction

Used for nearly all of literature's history as a narrative device, dreams generate psychological insight or turmoil not only in the fictional characters who experience them, but also in the readers engaging with dream-narratives [1]. Their uses in literature are often reflective of the prevailing theories of dreams in surrounding culture, but as it stands they have yet to be distinguished from the larger literary category of fictionality. In lieu of Andrew Piper's computational work on fictionality, we were surprised to see that the linguistic and cognitive components which made the literary category distinguishable from non-fiction were incredibly reminiscent of those which are employed to report dreams [2]. This prompted the following question : how distinguishable are dreams in fiction from other narratives?

Figure 1 : Alice In Wonderland Is A Quintessential Example Of Dream Narrative Being Embedded Within Fiction.

To answer this we intend to use the view of "dream-as-narrative-event," coined by J. Stephen Russell in his work on dream visions expressed in English poetry of the Middle Ages. He contends that the motif's narrative uses rely on both prevailing cultural conceptions of the dream as well

as its deeply personal, psychological effects [1].. Our work draws from the assertion made by Richard Walsh that dreams have significance for the field of narratology, in that their reportage is a distinct narrative act whose further study may offer "new insights into the nature of narrative fictionality and its effective power" [3]. Dreams and visions are listed prominently as examples of embedded and interruptive narratives which occur within a larger narrative across a literary work in Eisenberg and Finlayson's guidelines for annotation of narrative boundaries.

II. Related Work

Initial work on our classifier was largely informed by Razavi et al.'s dream sentiment analysis project, primarily by using a modified bag-of-words model which also employs part-of-speech tagging to preserve local word relationships in our training process [4]. Our goal of classifying dream-narrative from original narrative is possible mostly in part by classification work by Hendrickx et al. which concluded specific time and conversational expressions, descriptions of scenes and exhibiting a "lower discourse coherence than other personal narratives," are all strong distinguishing features of dreams from other personal narrative [5]. We hope to extend this work to dreams in literary contexts. Our focus on dream text and dream-narratives also follows arguments made by Patricia Anne Kilroe, where she points out that the degree of narrativity among dreams varies with length and coherence, a distinction we hope computational methods may better hone in on, as we can easily normalize or featurize both properties [6]. In particular, we think that dream-narratives with a lack of coherence may benefit from computational work done by Hoyt Long and Richard Jean So on the classification of stream-of-consciousness writing from realism, as well as its dissemination through world literature [7]. Like stream of consciousness (SOC), dream-narratives are varied but have a discrete set of classifiable features and dream-narrative's discrete phenomenological features bear resemblance to those found by Andrew Piper to be indicative of fictionality, and we plan to better identify the boundary between fiction and dream-narrative through the use of similar classifying features like the LIWC dictionary [2].

	Dream Data	Fiction Data
Source	Dreambank	HathiTrust
Number of entities (dreams or pages)	26342	134500 (from 500 volumes)
Avg number of tokens per page	121	255

Table 1. Data Summary Statistics

III. Materials and Methods

A. Data Collection

Our data at large consists of three datasets: dreams, fiction, and dreams reports found within fiction.

a. Dream Data Description :

For text-based dream reports, we used the UC Santa Cruz DreamBank dataset, a collection of over twenty thousand reported dreams gathered since the 1990s. These dreams are sourced from research

participants ranging from the ages of 7 to 74, and vary in length, complexity, and coherence.

b. Non-Dream Data :

For the non-dream fiction texts, the data is extracted from the HathiTrust Extracted Features dataset we used in class projects ; we've sampled 500 works of fiction from Ted Underwood's metadata as the extracted features data provides page-level word counts for the works..

We initially intended to use both fiction and nonfiction data in conjunction with the dream data but we eventually narrowed our scope to classifying dream narratives from fiction literature only as there was little to no overlap between the dream narratives and non-fiction narratives like Wikipedia, on which we tested our models during the pilot experiments.

B. Data Wrangling:

For the dream dataset, we downloaded around 26343 dreams from josauder git repo and tokenized them. Each dream was treated as a page and on an average each page had around 121 tokens. We organized these dreams and tokens into a dataframe adding additional features like count of each token within a given page(dream) and its part of speech as defined by the nltk library.

	page	tokens	counts	pos
1	0	my	4.0	PRP$
3	0	,	10.0	,
5	0	first	1.0	RB
6	0	thing	1.0	NN
7	0	i	12.0	NN
...	...	...	...	...
2478087	26342	celine	2.0	NN
2478088	26342	dion	1.0	NN
2478089	26342	sprinkle	1.0	NN
2478090	26342	oceanside	1.0	NN
2478091	26342	transient	1.0	NN

1772430 rows × 4 columns

Figure 2 : Dream Dataframe Preview

To ensure we don't bias our model or add any noise to it during training, we removed stopwords using the NLTK stopwords list and added custom stopwords which were explicit markers of dream narratives like 'dream', 'awake', 'sleep' etc.

For the fiction dataset, we processed the data by individual book volume rather than storing it into a dataframe due to processing constraints. We processed 500 fiction books and found on an average there were 253 tokens per page. Hathitrust API provides the unigram counts per page and parts of speech for each token.

Our validation data consists of 63 embedded dream narratives found within works of English fiction contained in the HathiTrust Digital Library. Our definition of dream narrative was based on the format of most dreams found in the DreamBank, that is those which resemble a direct dream report

from a specific character found in these volumes. Dream reports were queried by searching for the phrases "had a dream", "dream," "dreamt," and "vision". We used these tight constraints as a starting point for future work involving embedded dream narrative, which we hope will extend to subjects surrealist fiction and styles like stream-of-consciousness like that found in beat poetry, and also previously analyzed by Long and So [7].

We however noticed a huge variance in the maximum (~3000) and minimum(~10) count of tokens per page in the dream and fiction dataset after data processing and thus normalized the data by rescaling. We rescaled both the datasets to the range of 200 tokens per page by dividing the token counts in each page by the average number of tokens per page in the book and then multiplying the result by 200. In order to reduce deviation we then discarded all those pages which either had less than 30 or more than 300 tokens. This helped us to ignore pages with extremely low or high token counts and normalise all the counts to the same range.

We used 400 fiction books for training and testing with the Dream Bank dataset while 100 fiction books were exclusively used to test with hand-annotated validation data.

C. Topic Modeling

Before we went into classification, we ran a topic model against our dream dataset to understand what basic themes or archetypes we can expect to be favored by our model. The full results of topic modeling can be found in the Appendix but on a higher level here are the topics people mostly dream of : people, animals , places , objects , actions and uncertainty.

D. Feature Engineering

In the first stage of our experiments, we extracted certain features from our tokenized datasets and used a basic Logistic Regression Classifier to test their efficacy and also understand how each set of features was contributing towards identifying a dream narrative.

The features we created represent various textual markers of dream narratives. The features we eventually used were selected after a series of experiments. In each iteration of the experiment we added a single feature to our base Logistic Regression classifier and assessed the accuracy of our model. We also used the feature importance of Random Forest to find the dominant features.

Here is a description of the features used during the modeling process:

a. Unigram Features

We extracted word unigram features as even after removing explicit and obvious markers of dream texts there are certain words that can be highly characteristic of dream narratives and can be good independent predictors of dream text. Our ability to use n-gram features was limited by the availability of tokens only from the Hathi metadata and has been discussed in the Limitations and Future Work section.

b. LIWC Features

Linguistic Inquiry and Word Count or LIWC features are highly relevant to our analysis as they allow us to evaluate the lexical semantics of dreams

narratives to understand their emotional, cognitive, and structural components. A full description of the LIWC features can be found here.

c. Parts of Speech

Most people often dream about beings, things or places et al and while describing their dreams discuss these emotions and feelings w.r.t these entities. Such textual markers can be identified as grammatical descriptors or parts of speech. We used the POS tagger from the NLTK library as it processes a sequence of words and outputs specific tags for each word. A full description of the tags can be found here.

Some of the features we built but did not use in our modeling process were pronouns, punctuation, and word-length. We tried utilizing these features, but found that Unigram Counts, LIWC, and Parts-of-Speech were the most dominant features. Adding other features didn't improve the model accuracy. The reason could be that some redundant features like pronouns were already incorporated in the liwc feature.

E. Modeling

Before we began modeling, we split our data to hold out 20% of the data for testing with the validation dream set and the rest was used for training & testing with the Dream Bank data. We then compared the performance across five different classification algorithms:

a. Logistic Regression (with L1 and L2 Regularization)

b. Support Vector Classifier (SVC)

c. Multinomial Naive Bayes Classifier (MNB)

d. Random Forest Classifier (with varied depths and number of trees)

e. Perceptron Network

The models were chosen based on its compatibility and performance with the sparse features.

F. Evaluation.

Metrics : We rated the effectiveness of these models by calculating their accuracy as well as the F1 score for each model using the sklearn library. The F1 score is a weighted average of the precision and recall, wherein an F1 score of 1 is the best and a score of 0 is the lowest. We used the F1 score as both false positives and false negatives were essential to determine the effectiveness of our classifiers and the relative contribution of precision and recall to the F1 score are equal.

Validation : In order to ensure that there is no overfitting happening we performed two level validation testing.

First, we evaluated our models on a validation data set which exclusively consisted of dreams from the dream bank dataset not used in training. Herein we evaluated the efficiency of our models to predict a dream correctly from the dream bank dataset.

Secondly, to test the efficiency of our model in identifying dreams within fiction (the goal of this project), we built a hand annotated validation set of around 63 dreams found in fictional texts. We manually built this dataset from HathiTrust by doing advanced search for fictional texts which are dream narratives. Each of these hand annotated dreams-within-fiction texts consisted of an average 250 tokens (similar to the training set) and were derived from books of English fiction. The model

performance was also tested with this set of dreams and random selection of fiction books not used in previous steps.

Confidence Intervals : To test the margin of error for our models, we ran two tailed t-tests to determine the confidence interval for the predictions made. We randomly sampled from our held out set of 100 works of fiction and confirmed that there is a 95% likelihood that the range [0.777,0.850] covers the true model accuracy.

IV. Results and Discussion

Model	Metrics	
	Accuracy	F1 Score
Logistic Regression with L2 regularization	99.1%	98.6%
RandomForestClassifier (max_depth=10, n_estimators=200)	97.5%	96.09 %
SVC	98.9%	98.35 %
MNB	68.8%	6.40%
Perceptron	98.8%	98.08 %

Table 2: DreamBank dream data validation set metrics.

Model	Metrics	
	Accuracy	F1 Score
Logistic Regression with L2 regularization	81.8%	47.6%
RandomForestClassifier (max_depth=10, n_estimators=100)	86.84%	64%
SVC	75.9%	49.72 %
MNB	79.9%	0%
Perceptron	45.58%	38.25 %

Table 3 : Hand annotated dream-within-fiction validation data set metrics

On the hand annotated validation data set our best model of all was the Random Forest classifier . It produced an impressive accuracy of 86.84% along with a F1 score of 64%.

But in terms of interpretability of results , the Random Forest classifier is a double-edged sword. While it is largely effective as a nonlinear model, the Random Forest model is unable to identify feature importance based on classes/polarity and hence doesn't allow us to understand which features are more predictive of dreams and which are for fictional narratives.

The Logistic Regression and Linear Support Vector (SVC) classifiers had the highest explainability of all the models we tried as they provide weights of the features that the classifier used internally to make its predictions. It also gave an accuracy of

81.8% with a F1 score of 47.6% on the hand annotated validation dataset.

Our unigram feature indicates that individual words which predictably denote dreams include those indicative of first-person actions and interactions with the surrounding world. Words like said, looked, came, walked, showed, and went are also all in the past-tense, which is something to be expected considering dreams are often recounted after the experience of having them. Taken in concert with the words remember, wanted, and realize it's clear that our classifier is picking up on the fact that dream reports are constructs of memory with a certain degree of fallibility and uncertainty, often reflective of an actor entering a context which they selectively become aware of. This can lead to some details being favored greater than others, and the presence of unigrams which indicate specific desires and needs within a dream-state are in support of existing dream theories like Freudian wish-fulfillment [10].

Words which denote fiction depict a subject's greater awareness of their surrounding environment and the perceptual, physiological, and cognitive processes which grant that awareness. Words such as world, earth, mind, eyes, moment, form, indeed, certain, breath, and silence all reflect a greater degree of perceptual certainty and descriptive richness employed by any given narrator than those found in dreams. This falls in line with Hendrickx et al.'s assertion that dreams exhibit a lower discourse coherence than other narrative [5].

The LIWC categories of Cognitive Processes, Time Orientation Words including both those which focus on past and present, the presence of negative emotion, as well as the use of the word I are overwhelming predictors of dream narrative. The single category of "I" was the most heavily weighted feature, which logically follows our constraints on the first-person dream narrative. In relation to unigrams, cognitive processes seen to fall within the LIWC categories are simpler than that of fiction, often representing basic perceptual categories like "see" or "know", further supporting the notion that dream reports have a lower degree of coherence not just for those readings, but for the subjects who experienced them.

LIWC categories indicative of fiction are the pronouns of he and she, as well as the word you. These second and third-person categories are in contrast to the strong first-person character of dream reports, and likely represent a more complex fictional narrative composed of many more instances of dialogue and reference to external characters. Reaffirming our claim that uncertainty is a strong indicator of dream narrative as shown by unigrams, the LIWC category of certainty is a strong indicator of non-dream fiction, again highlighting the non-ambiguity and richer understanding of a character or narrator's waking world.

The parts of speech features isolated from our data are also in agreement with our findings thus far. Past-tense verbs and personal pronouns are strong indicators of distinguishing between fictional narrative and dream narrative. Adverbs are favored towards dream narratives, likely describing the many non-cognitive actions found by our unigram feature. Unsurprisingly, both singular and plural proper nouns are indicative of fictional narrative, reflecting a richer understanding of the surrounding world in the form of identifiable characters, places, or other entities.

V. Conclusion

Our results show that dreams-within-fiction can be predicted with high accuracy using simple features like unigrams, LIWC and parts of speech. We found that dream narratives-within-fiction are distinguishable from other narratives as they are mostly in past tense and demonstrate a more basic sense of cognition in the form of sense-perception. Altogether, dreams are narrated with a higher level of uncertainty than fiction, in which we found features to be much more descriptive of setting and reflective of a greater understanding of the narrator's surrounding world.

VI. Limitations and future work

One of the key limitations of this work was that we could develop only unigram features and not n-gram features on the dream dataset. This was because we only had unigram tokens available from HathiTrust (and not full text due to administrative issues) for the fiction data and hence had to restrict the dream data to unigram features as well.We would also like to add more varied dream text data to our corpus.

Finally, we would have liked to assimilate more data in the validation dataset for us to eventually test the efficiency of our models. The current validation dataset had to be hand annotated which is a laborious method and yields low output over a large time.

Analysing the current results, we are highly optimistic about the future prospects of this project and believe that further improvements can be made as follows:

1. We can process ngram features if fictional texts are made available to us and not just the metadata. This will allow us to create more meaningful features and also to use more sophisticated models like LSTM and Bert to achieve better results.
2. Model efficiency can be improved and we can also do close readings if annotated dream-within-fiction data is available.
3. With more data available, we plan to draw a temporal plot of dream narrative in literature with a 95% confidence interval over a period of 200 years. We can run our classifier through the literary texts in a sliding window fashion and make a soft prediction on its dream quotient.

VII. References

[1] Russel, Stephen J. The English Dream Vision https://kb.osu.edu/bitstream/handle/1811/24692/1/THE_ENGLISH_DREAM_VISION.pdf

[2] Andrew Piper, "Fictionality", Cultural Analytics Dec.20, 2016. DOI: 10.22148/16.011 https://culturalanalytics.org/article/11067-fictionality

[3] Walsh, Richard. "Dreaming and Narration" https://www.lhn.uni-hamburg.de/node/70.html

[4] Matwin Stan, De Koninck Joseph, Razavi Amir H., and Amini Ray Reza. "Classification of Dreams Using Machine Learning." JB. Frontiers in Artificial Intelligence and Applications 215, no. ECAI 2010 (2010): 169–74. https://doi.org/10.3233/978-1-60750-606-5-169

[5] Iris Hendrickx, Louis Onrust, Florian Kunneman, Ali Hürriyetoğlu, Wessel Stoop, and Antal van den Bosch, "Unraveling reported dreams with text analytics," Digital Humanities Quarterly. 2017. arXiv:1906.05433 https://arxiv.org/pdf/1612.03659.pdf

[6] Kilroe, Patricia A. "The Dream as Text, the Dream as Narrative." Dreaming 10, no. 3 (2000): 125–37. https://doi.org/10.1023/a:1009456906277

[7] Long, Hoyt, and Richard Jean So. "Turbulent Flow." Modern Language Quarterly 77, no. 3 (August 15, 2016): 345–67. https://doi.org/10.1215/00267929-3570656

[8] Schreiber, Evelyn Jaffe. "Dream Visions and Stream-of-Consciousness: The Conscious and Unconscious Search for Meaning." *Journal of the Fantastic in the Arts* 7, no. 4 (28) (1996): 4-15. Accessed November 3, 2020. http://www.jstor.org/stable/43308265.

[9] Freud, Sigmund. *The Interpretation of Dreams.* Trans. And Ed. James Strachey. NY: Avon, 1965

VIII. Appendix

1. Complete Results for all models

   https://docs.google.com/spreadsheets/d/1AI80awUXMOQIIw7rIs9IJ0OkTt6Xj5EH8zN_FCRvVls/edit?usp=sharing

2. Topic Modeling Results

   https://drive.google.com/file/d/1_iAIn5lmqNePwkmB4gnf4ktD_YD_BeSD/view?usp=sharing

3. Hand Annotated Validation Dreams Dataset

   https://drive.google.com/file/d/1IoiuCZInaM5isvWkvjFtv8Mfnc6jScv-/view?usp=sharing