Gender Detection and Stylistic Differences and Similarities between Males and Females in a Dream Tales Blog

Raffaele Manna
UNIOR NLP Research Group
University L'Orientale
Naples, Italy
rmanna@unior.it

Antonio Pascucci
UNIOR NLP Research Group
University L'Orientale
Naples, Italy
apascucci@unior.it

Johanna Monti
UNIOR NLP Research Group
University L'Orientale
Naples, Italy
jmonti@unior.it

Abstract

English. In this paper we present the results of a gender detection experiment carried out on a corpus we built downloading dream tales from a blog. We also highlight stylistic differences and similarities concerning lexical choices between men and women. In order to carry the experiment we built a feed-forward neural network with traditional sparse n-hot encoding using the Keras open source library.

1 Introduction

It is generally accepted that dreams are just an unconscious production, and that represent a type of non-manipulable happening. However, many people believe that dreams are premonitory of future events as well as representations and reworkings of past events. Humans tend to preserve all personal events, some of them in the form of a diary, namely the best method to tell an event and keep its aura of magic.

Until recently, dream reports were relegated to the the pages of paper journals or revealed to familiar people. At an earlier time, dreams are gathered from sleep research labs, psycho-therapeutic and in patient settings, personal dream journals and occasionally classroom settings where "most recent dreams" and "most vivid dreams" are collected as in (Domhoff, 2003).

Social media have opened millions of pages where people feel at ease to confess their thoughts, their experience and even their secret fantasies. These platforms such as Twitter, Facebook and web blogs are a good ground for computational text analysis research in social science and mental health assessment via language.

Diary narratives represent a field already investigated by researchers. The recent development of web communities focused on telling dreams allows researchers to access and discover new characteristics related to the language of dreams. Stylistic and linguistic features of dreams in blog reports are essential in order to detect writing style and content differences between men and women, but also enable future researches associated to the different types of personality and styles associated with mental health diagnoses and therapeutic outcomes.

The aim of this paper is to show that despite dreams are just an unconscious production, there are several stylistic differences between the reports of dreams by males and females on online blogs. The model we built is able to represent and classify all stylistic differences.

Moreover, this research represents a preliminary step in the field of dream tales which will be followed by an attempt to find stylistic differences between dream tales and other forms of self narration (i.e. travel tales).

The paper is organized as follows: in Section 2 we introduce Related Work, in Section 3 we describe the corpus we built and the blog. Methodology is described in Section 4 and Results are in Section 5. In Section 6 we present our Conclusions and we introduce Future Work.

2 Related Work

Textual analysis of dream reports is still not a completely investigated field in NLP. One of the purposes of computational dream report analysis lies in understanding how and why a dream narrative differs from a waking narrative (Hendrickx et al., 2016). For example, if a dream description contains more function words than a waking narrative, what is the relationship between the content of dreams and the use of more function words?

Earlier studies were conducted by (Domhoff, 2003 and Bulkeley, 2009). In their researches, dream reports are analyzed and a systematic category list of words that can be used for queries and word-frequency counts in the DreamBank.net is provided. The categories are related to the content of dreams and used to retrieve the mentions of emotions, characters, perception, movement and socio-cultural background.

On the basis of this approach (Bulkley, 2014) update the categories list and evaluate it on four datasets of the *DreamBank* corpus. It has been shown that this type of word analysis can be applied to detect the topics of dreams. In addition, this latter contribution provides evidence that it is possible to guess about a person's life and activities, personal concerns and interests based on an individual dream collection .

Other works focus on identifying the emotions in the reports of dreams. In particular (Razav et al., 2014) use a machine learning method to assign emotion labels to dreams on a four-level negative/positive sentiment scale. In their research, dreams are represented as word vectors and dynamic features are included to represent sentiment changes in dream descriptions.

In a more accurate sentiment analysis, (Frantova and Bergler, 2009) train a classifier, based on semi-automatically compiled emotion word dictionaries, in order to assign five fuzzy-emotion categories to dream reports. Then, they compare their results against a sample from the *DreamBank* that is manually labeled with emotion annotations.

In some non-computational studies and aimed at highlighting gender differences (Schredl, 2005; Schredl, 2010), dream reports are used to spot gender differences in dream recall. The first research demonstrates that gender differences in dream recalls and dream contents are stable. Human judges are able to correctly match the dreamer's gender based on a single dream report with a probability better than chance. Based on these findings, in the latter study the stability of gender differences in dream content is analyzed over time. Two dream themes (work-related dreams and dreams of deceased persons) were investigated and gender differences resulted quite stable over time. In (Mathes, 2013) gender differences are associated to personality traits. The analysis indicate that some of the big five personality dimensions might be linked with some dream characteristics such as characters and the occurrence of weapons or clothes in dreams.

In psychiatric studies, the gender variable is identified as a predictive for psychotic behaviors and disorders. In (Thorup, et al., 2007), the authors showed that, in psychotic patients, the gender-related variable has a role in showing different psycho-pathological characteristics and different social functioning. Although no dream samples were taken as a subject in this study.

Dream diaries refine the research in uncovering connections between dreams and dreamer's socio-cultural background, mental conditions and neuro-physiological factors. The language of online dreams in relation to mental health conditions has yet to be analyzed, however prior laboratory research suggests that dream content may differ according to clinical conditions.

In (Skancke et al., 2014), emotional tone, themes and actor focus in dream report were associated with anxiety disorders, schizophrenia, personality and eating disorders. However, it is not clear whether dream content can be predictive with respect to mental disorders.

In (Scarone, 2008), the hypothesis of the dreaming brain as a neurobiological model for psychosis is tested by focusing on cognitive bizarreness, a distinctive property of the dreaming mental state defined by discontinuities and incongruities in the dream report, thoughts and feelings. Cognitive bizarreness is measured in written reports of dreams and in verbal reports of waking fantasies in thirty schizophrenics and thirty normal controls. The differences between these two groups indicate that, under experimental conditions, the waking cognition of schizophrenic subjects shares a common degree of formal cognitive bizarreness with dream reports of both normal controls and schizophrenics. These results support the hypothesis that dreaming brain could be a useful experimental model for psychosis. Taking advantage of all the above considerations and mixing the psychiatric and neurobiological information of the studies shown, the present research wants first of all to reveal the differences between genders in dreams. And as a future goal, starting from the hypothesis of cognitive similarity between dreams and psychoses and using dreams as an experimental path, to clarify the relationship between gender and psychosis.

3 Dataset Description

The web is full of blogs, where people can share opinions, questions and personal feelings and thoughts about their own life. Furthermore, people also share their dreams, one of the most personal hidden aspects of life.
It is very easy to find a blog in which thousands of people share their “dream experiences”, sometimes discovering that other people have had similar experiences dictated by similar life styles.
We investigated a blog, called *SogniLucidi*, on which every day thousands of people tell their dreams and nightmares, mixing their nightly fantasies with their unconscious writing style choices. *SogniLucidi*, that literally can be translated in *LucidDreams* took its name from a term coined by the Dutch psychiatrist Frederik van Eeden in 1913: it describes the situation in which dreamers are aware that they are dreaming.
There are many techniques that, when correctly applied, allow dreamers to obtain a “Lucid Dream” and that we report for completeness: **CAT** (*Cycle Adjustment Technique*), **MILD** (*Mnemonic Induction of Lucid Dreaming*), **WBTB** (*Wake Back To Bed*), **WILD** (*Wake Initiated Lucid Dreams*), **RCT** (*Reality Control Test*) and **ITES** (*Induction Through External Stimulus*).
The corpus we built for the investigation is balanced with gender and the number of authors analyzed is not randomly selected but represents the precise number of participants to the blog.

3.1 Dataset Statistics

In this paragraph, we present the resulting statistics obtained using the NLTK module together with other statistics formulas for the analysis of the corpus we built on *SogniLucidi* blog. In Table 1 we report two important statistics about words: the number of tokens in texts written by men and women and word types. We can notice that there is a big difference in the number of tokens used by Males (80629) and Females (57673).

	Males	**Females**
Number of Tokens	80629	57673
Word Types	12254	11158

Table 1: Words’ statistics in the whole corpus in terms of Number of Tokens and Word Types.

In Tables 2 and 3 we present four lists of six exclusive nouns and six exclusive verbs used by men or women. Both exclusive nouns and exclusive verbs are the most relevant for frequency for Males and Females classes. Verbs are reported in their base form. The results indicate, without interpretative effort for a human, that most relevant topics given these high frequency words are associated to activities and events that the dreamers want to happen, in settings and adventurous situations for male dreamers. Meanwhile dreamers belonging to Females class seem to set their dreams in a baleful scenario, where “transizione” (transition) and “trapasso” (transition) mean that they dream about twilight state, beyond death or they fantasize about surreal activities.

Males	**Females**
destinazione (destination)	balzo (bound)
esplosione (explosion)	luce (light)
foresta (wood)	nuvola (cloud)
lenzuola (linens)	piscina (swimming pool)
spiaggia (beach)	transizione (transition)
terrazze (terraces)	trapasso (transition)

Table 2: Most frequent Exclusive Nouns in the whole corpus.

Males	**Females**
assomigliare(to resemble)	affrontare(to face)
baciare(to kiss)	cadere(to fall)
funzionare(to function)	ragionare(to reason)
ottenere(to obtain)	stringere(to tighten)
scomparire(to disappear)	succedere(to happen)
superare(to overcome)	volare(to fly)

Table 3: Most frequent Exclusive Verbs in the whole corpus.

Lastly, in Table 4 we report the average of tokens per sentence.

Males Tokens AVG	**Females Tokens AVG**
18,74 tokens/sentence	10,01 tokens/sentence

Table 4: Average of tokens per sentence in texts written by men and women.

4 Methodology

The training corpus consists in dream text descriptions written by two groups of authors:

- 28 Male authors;
- 28 Female authors.

The corpus is balanced and labelled with gender. Gender annotation has been done manually and based on the name of the users, their profile photos and description. For each author, a total of fifteen texts about dreams are provided. Authors are coded with an alpha-numeric author-ID. For each author, the last fifteen texts about dreams have been retrieved from the personal web diary's timeline. As a result, the time frame of the dream reports might vary from days to months, depending on how frequently users report their dreams on the blog. To train our classification model, we exploited the descriptions of dreams only and not the comments (both comments of the authors and comments of other members of the *SogniLucidi* blog).

4.1 Preprocessing

For preprocessing we used the Python library *BeautifulSoup* along with same regex procedures. We performed the following preprocessing steps:

- Removing the html tags;
- Removing URLs;
- Removing @username mentions;
- Lower-casing the characters;
- Detecting stop-words by document frequency and removing. Only n-grams that occurred in all documents has been considered a stop-word and ignored.

4.2 Features

Feature selection is a very critical step in any model. For feature selection we use the sklearn utilities *SelectKbest*. It selects the n-best feature based on a given criterion. In our experiments, the features are selected on the *f_classif* criteria. This function perform an *ANOVA* test, a type of hypothesis test, on each feature on its own and assign that feature a p-value. The SelectKbest rank the features by that p-value and keep only the n-best features. The feature set for the dream dataset benefits from word trigrams in addition to other n-grams. In our final model, we use the following n-grams features: Word unigrams, bigrams and trigrams.

Word level n-grams used the following parameters:

- Minimum document frequency = 2. Terms with a document frequency lower than would be ignored;
- Term frequency-inverse document frequency (tf-idf) weighting;
- Maximum document frequency = 1.0 or rather terms that occur in all documents would be ignored.

4.2.1 Classification Model

We built a neural network to perform the gender detection issue. We decided to run a feed-forward neural network with traditional sparse one-hot encoding with the *Keras* open source library. After a parameters selection, the model obtained the best performance with an Adam optimizer and a learning rate of 0.32, feeding it with a batch size of seventy and training for thirty epochs. Moreover, the input layer of sixty-five neurons with an initialization using a norm kernel. Then, a RELU activation function was applied, followed by a dropout layer. During optimization, we found that a relatively big dropout rate of 0.5 outperformed the smaller dropout rates. The output layer is a single neuron, followed by a linear activation function. The feature set provided to the model was an n-hot encoding of the uni-, bi- and trigrams.

5 Results

In this section we describe the results on the training data and the test data. The data we used was split into training and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test set (or subset) in order to test our model+ prediction on this subset. We calculated accuracy scores on the training data, both on validation set (Dev set) of 0.3 and Test set of 0.2. The performances (both for Dev test and Test set) are shown in Table 5 in terms of Accuracy, Precision and F1 Score. We obtained roughly the same results for Accuracy in Dev set and the Test set, 0.794 and 0.775, respectively.

Finally, in order to compare our approach, we considered two other baseline models namely Multinomial Naive Bayes (MNB) and Linear Support Vector Machine (SVM) besides the feed-forward

	Dev set	Test set
Accuracy	0.796	0.776
Precision	0.937	0.917
F1 Score	0.803	0.786

Table 5: Performances in Dev set and Test set in terms of Accuracy, Precision and F1 Score.

neural network for performance comparisons on Test set.

MNB	SVM
0.411	0.588

Table 6: Baseline Accuracy Comparisons.

To assess the performance of the model, the Root Mean Square Error (RMSE) was computed. RSME measures the distance of the predicted value to the true value. It is a measure of error, so the lower is the score, the better is the performance. We show RMSE results in Table 7.

Dev set	Test set
0.233	0.224

Table 7: RMSE of the feed-forward model on the Dev set and when using Test set.

Using classification accuracy alone when evaluating the performance of the classification algorithm could be misleading, especially if the dataset- as in our case - is limited in size or is unbalanced or contains more than two classes. Hence, a confusion matrix is used to evaluate the results of the experiments. The confusion matrix *M* is a *N*- dimensional matrix, where *N* is the number of classes, that summarizes the classification performance of a classifier with respect to Test set and Dev set, both as in our case. Each column of the matrix represents predicted classifications and each row represents actual defined classifications. As shown in Table 8, during the validation phase, the classifier made a total of two hundred-sixteen predictions, while during the test phase the classifier made a total of two hundred-fourteen predictions. Out of two hundred-sixteen cases in validation, the classifier predicted “Females” forty-four times and sixty-four “Males”. Actually, sixty people in the sample belong to “Females” class and forty-eight to “Males” class.

	Males	Females
Males	45	3
Females	19	41

Table 8: Confusion Matrix on Dev set.

After this intermediate phase and after having tuned the parameters in order to optimize the model on the previous results, the classifier made a total of two hundred-fourteen predictions during the test phase. Out of two hundred-fourteen predictions, the model predicted “Females” forty-three times and sixty-four “Males”. Indeed, fifty-nine people belong to “Females” class and, as predicted during the validation phase, forty-eight to “Males” class. We report gender prediction results on test data in the confusion matrix in Table 9.

	Males	Females
Males	44	4
Females	20	39

Table 9: Confusion Matrix on Test set.

6 Conclusions and Future Work

In this paper we have shown our results on gender detection in dream diaries and writing styles differences and similarities between males and females in dream tales. First we explored the vocabulary of dream descriptions for both the genre-class by listing some of the representative words for each genre. Then, we evaluated our gender detection model on the dream reports dataset. The model succeeded in obtaining good results managing to distinguish a good part of dreams made by men or women. This research represents our preliminary step in the field, toward subsequent studies, in which we are trying to detect stylistic differences between dream tales and personal descriptive narratives, such as travel tales and other forms of self-narration.

Acknowledgments

This project has been partially supported by the PON Ricerca e Innovazione 2014/20 and the POR Campania FSE 2014/2020 funds.

References

Altszyler, E., Sigman, M., Ribeiro, S., Slezak, D. F.. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520.*

Altszyler, E., Ribeiro, S., Sigman, M., Slezak, D. F.. (2017). The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text. *Consciousness and cognition, 56, 178-187.*

Bulkeley, K.. (2009). Seeking patterns in dream content: A systematic approach to word searches. *Consciousness and cognition, 18(4), 905-916.*

Bulkeley, K.. (2014). Digital dream analysis: A revised method. *Consciousness and cognition, 29, 159-170.*

Coelho, H.. (2010). Classification of dreams using machine learning. In *ECAI: 19th European Conference on Artificial Intelligence: Including Prestigious Applications of Artificial Intelligence (PAIS-2010): Proceedings (Vol. 215, p. 169).*

Domhoff, G. W.. (2003). *The scientific study of dreams: Neural networks, cognitive development, and content analysis.* American Psychological Association.

Domhoff, G. W., Schneider, A.. (2008). Similarities and differences in dream content at the cross-cultural, gender, and individual levels. *Consciousness and cognition, 17(4), 1257-1265.*

Frantova, E., Bergler, S.. (2009). Automatic emotion annotation of dream diaries. In *Proceedings of the analyzing social media to represent collective knowledge workshop at K-CAP 2009, The fifth international conference on knowledge capture.*

Hawkins, I. I., Raymond, C., Boyd, R. L. 2017. *Such stuff as dreams are made on: Dream language, LIWC norms, and personality correlates*, Dreaming, 27(2), 102.

Hendrickx, I., Onrust, L., Kunneman, F., Hrriyetolu, A., Bosch, A. V. D., Stoop, W. 2016. Unraveling reported dreams with text analytics. *arXiv preprint arXiv:1612.03659.*

Koppel, M., Argamon, S., Shimoni, A. R., 2002. *Automatically categorizing written texts by author gender. Literary and linguistic computing*, 17(4), 401-412.

Mathes, J., Schredl, M.. (2013). Gender differences in dream content: Are they related to personality?. *International Journal of Dream Research.*

Mechti, S., Jaoua, M., Belguith, L. H., Faiz, R., 2013. *Author profiling using style-based features*, Notebook Papers of CLEF2.

McNamara, P., Duffy-Deno, K., Marsh, T.. (2019). Dream content analysis using Artificial Intelligence. *International Journal of Dream Research, 42-52.*

Mukherjee, A., Liu, B.. (2010, October). Improving gender classification of blog authors. In *In Proceedings of the 2010 conference on Empirical Methods in natural Language Processing (pp. 207-217).* Association for Computational Linguistics.

Niederhoffer, K., Schler, J., Crutchley, P., Loveys, K., Coppersmith, G.. (2017, August). In your wildest dreams: the language and psychological features of dreams. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical PsychologyFrom Linguistic Signal to Clinical Reality (pp. 13-25).*

Nielsen, T. A., Stenstrom, P.. (2005). What are the memory sources of dreaming?. *Nature, 437(7063), 1286.*

Rangel, F., Rosso, P.. 2013. *Use of language and author profiling: Identification of gender and age*, Natural Language Processing and Cognitive Science, 177.

Razavi, A. H., Matwin, S., De Koninck, J., Amini, R. R.. (2014). *Dream sentiment analysis using second order soft co-occurrences (SOSCO) and time course representations.* Journal of Intelligent Information Systems, 42(3), 393-413.

Scarone, S., Manzone, M. L., Gambini, O., Kantzas, I., Limosani, I., D'agostino, A., Hobson, J. A.. (2008). The dream as a model for psychosis: an experimental approach using bizarreness as a cognitive marker. *Schizophrenia Bulletin, 34(3), 515-522.*

Scarpelli, S., Bartolacci, C., D'Atri, A., Gorgoni, M., De Gennaro, L.. (2019). The functional role of dreaming in emotional processes. *Frontiers in Psychology, 10.*

Schredl, M., Sahin, V., Schfer, G.. (1998). Gender differences in dreams: do they reflect gender differences in waking life?. *Personality and Individual Differences, 25(3), 433-442.*

Schredl, M., Ciric, P., Gtz, S., Wittmann, L.. (2004). Typical dreams: stability and gender differences. *The journal of psychology, 138(6), 485-494.*

Schredl, M., Piel, E.. (2005). Gender differences in dreaming: Are they stable over time?. *Personality and Individual Differences, 39(2), 309-316.*

Schredl, M., Becker, K., Feldmann, E.. (2010). Predicting the dreamers gender from a single dream report: A matching study in a non-student sample. *International Journal of Dream Research.*

Schredl, M., Noveski, A.. (2018). Lucid Dreaming: A Diary Study. *Imagination, Cognition and Personality, 38(1), 517.* https://doi.org/10.1177/0276236617742622

Siclari, F., et al. (2017). The neural correlates of dreaming. *Nature neuroscience, 20(6), 872.*

Silberman, Y., Bentin, S., Miikkulainen, R.. (2007). Semantic Boost on Episodic Associations: An EmpiricallyBased Computational Model. *Cognitive Science, 31(4), 645-671.*

Skancke, J. F., Holsen, I., Schredl, M.. (2014). Continuity between waking life and dreams of psychiatric patients: a review and discussion of the implications for dream research.*International Journal of Dream Research.*

Thorup, Anne and Petersen, Lone and Jeppesen, Pia and Ohlenschlæger, Johan and Christensen, Torben and Krarup, Gertrud and Jorgensen, Per and Nordentoft, Merete. (2007). Gender differences in young adults with first-episode schizophrenia spectrum disorders at baseline in the Danish OPUS study.*The Journal of nervous and mental disease, 195(5), 396-405*

Van Eeden, F.. (1913, July). A study of dreams. In *Proceedings of the Society for Psychical Research.Vol. 26, No. Part 47, pp. 431-461.*