(IJRSSH) 2023, Vol. No. 13, Issue No. IV, Oct-Dec **e-ISSN: 2249-4642 p-ISSN: 2454-4671**

Clustering Tiny Tales

***Ankita Nandy, **Anahit Tahmasyan Bal**
**Independent Researcher,*
Hooghly, West Bengal, India
*** Independent Researcher,*
Istanbul, Turkey

DOI:10.37648/ijrssh.v13i04.005

[1]*Received: 28 August 2023; Accepted: 15 October 2023; Published: 17 October 2023*

ABSTRACT

Social media gives its users an open space to express themselves, which becomes the fodder for numerous research works aimed at understanding the opinion, behavior, and attitudes of the users. While textual analysis was a domain reserved for human eyes, advancement in machine learning has made the analyses of the plethora of tweets, posts, reviews et cetera automated. Such texts often include an explicit mention or reference to the object of interest. On the other hand, creative pieces such as stories, often convey the thought in an implicit manner, and have witnessed limited experimentation with machine learning driven techniques. This work assembles a corpus of Terribly Tiny Tales, a social media presence publishing short stories, originally limited to 140 characters. A manual thematic analysis is followed by an attempt to obtain such relevant clusters through short text clustering techniques of Top2Vec and BERTopic, where the latter is found to generate more meaningful clusters. The quality and validity of a document cluster are based on human understandability; thus, some subjectivity is inherent, and there is plentiful room for further research.

Keywords: *Terribly Tiny Tales; topic modelling; text clustering; Top2Vec; BERTopic*

INTRODUCTION

Storytelling has been an art, well preserved and handed down over generations, for centuries, across all civilizations. Besides entertainment, narratives are sources of wisdom and have religious, sociocultural, and historical value. The tradition of storytelling has had an oral component, often enriched with dramatization, music, puppetry, dance etc. The written compilations, available as printed books, audiobooks et cetera are in wide circulation. Myths, legends, stories, anecdotes, and riddles, unified under the term "narrative" in the following text, defines how society is known or remembered and reflects what they hold important [15]. Their analyses are used to understand the people who composed, popularized, and preserved them.

Motifs can be understood as the key ideas that persist in the transmission of a narrative. Reference [15] explores the association of such in folklore and the dominant occupations, trades and socio-political structures of the societies preserving them. Popular beliefs are used to trace the evolution of rituals and customs, thus providing them with a historical context [11]. As narratives are not limited to content handed over through traditions, rather are in plenty in print and social media, in the form of books, blogs and social media posts et cetera, their analyses provide interesting insights into the minds of the authors as well as the audience which consumes the content. The volume of such narratives serves as fodder for research across myriad fields. In [12], the deep dissatisfaction of the middle class, portrayed through the protagonist of Dickens' David Copperfield, comes forth. This work also sources its content under study from social media. Though manual thematic analysis makes up a major chunk of such research works, textual analysis has witnessed rapid advancement with statistical and machine learning techniques, especially if a large digitized corpus can be

[1] How to cite the article: Nandy A., Bal A.T..; October 2023; Clustering Tiny Tales; *International Journal of Research in Social Sciences and Humanities*, Vol 13, Issue 4, 42-45, DOI: http://doi.org/10.37648/ijrssh.v13i04.005

accessed: stories can be labeled, and classified. The motifs that define the storyline can be identified through unsupervised techniques, and new narratives can be synthesized.

The task of extracting semantic information in text documents has been a challenge to the research community, and it had made progress on several related subtasks such as semantic similarity, and phrasal equivalence, through the announcement contests, as elaborated in [1][14]. While [6] discusses the extraction and classification of relations between entities in formal documents such as research papers, [9] presents the popular techniques in clustering informal texts, such as tweets, creative texts, such as short stories, remain vastly unexplored territory. The brevity of the content and diversity of the topics discussed in short texts present some unique challenges [2]; each text will have limited information and the vast vocabulary will yield a high dimensional sparse bag-of-words representation of the documents under study. In addition to these, the content under study being creative in nature, in contrast to the data considered by [20], rarely mentions the topic of interest in an explicit manner.

"Terribly Tiny Tales" came forth as a Facebook page in 2015. It featured stories, as short as 140 characters, later the limit was relaxed. The brevity of the content was the highlight, and led to several commercial collaborations. The content is available publicly in the form of images on platforms such as Facebook and Instagram et cetera. on a variety of subjects. This collection of 229 stories undergoes a basic thematic analysis, followed by the application of two techniques BERTopic and Top2Vec, to assess the meaningfulness of the clusters output by these techniques.

MATERIALS AND METHODS

A. Data

As the stories on TTT are uploaded as pictures, using Tesseract Optical Character Recognition (OCR) [19], the text from the images is retrieved and compiled into a CSV file for further usage. The Tesseract OCR identifies connected components as blobs, blobs into text lines. These text lines are broken down into words and letters which are mapped to their character using an adaptive classifier, which learns as it processes. The classifier runs through the page twice, so that the results of the first pass can be improved during the second pass. The technologies used in Tesseract were way ahead of its competitors. It made its impression in the 1995 UNLV annual test of OCR accuracy by surpassing other contenders by a wide margin. A total of 229 images were gathered. The data was cleaned to remove/substitute special characters and non-English words.

B. Topic Modelling

Reference [9] performs a comparative analysis of four topic modeling techniques on Twitter data. These techniques are Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Top2Vec and BERTopic. In contrast to the former two, the latter two need almost nil pre-processing, and the original text document is analyzed as is.

Representation of textual elements in some numeric format has been the motivation behind rudimentary models such as bag-of-words or word count/frequency-based feature generation. Word embeddings are defined by [3] as "dense, distributed, fixed-length word vectors, built using word co-occurrence statistics as per the distributional hypothesis." They represent words/sentences/documents using fixed-length numerical vectors, based on features such as co-occurrence and contextual similarity. The vector representations of similar entities will have higher similarity than dissimilar ones. Several pre-trained word embeddings are publicly available, one of them is Google's Universal Sentence Encoder (USE), which is employed in this work. USE generates a 512-element long vector for each text. The techniques employed in this work, Top2Vec [4] and BERTopic [10] both represent the documents under study using embeddings. These do not need any prior processing of the data, such as the removal of stop words, however this work dropped stop words to obtain better representation of the topics.

RESULTS AND DISCUSSION

Through thematic analysis, five primary themes were identified in the stories, elaborated as follows.

Romantic love. Love and its titillations dominate all forms of media, this being no exception. Trivial but memorable experiences bring forth the sheer excitement of being with loved ones, also exposing some social ills, "Love demands faith, trust, sacrifice and respect. And sometimes, it demands a brand-new BMW too."

Relationships. People today are virtually connected yet feel lonely [5][8][17]. A lot many stories are centered around the bonds we share with family members, or with random people. This might indicate an inherent loneliness in the

audience which the stories attempt to quench. This has been aptly worded in one of the stories, "Nothing defines loneliness in today's age more than trying to find companionship while flicking desperately through left swipes at 3AM."

Self-love. In a social media world chasing external validation, there are stories which remind the reader of the heroism that seeks no audience. Cherishing seemingly small achievements is a key theme in these stories, appearing in 17 of them. On example being, "Riding a bike may not be extraordinary, but failing off one and getting back on takes resilience. And that is a superpower of its kind."

Choice. As the audience on social media platforms is often younger than 20 years of age [18], exploring the choices one is presented with and arriving at a decision can be a daunting task. 75 of such stories are identified. Finding oneself at crossroads, a mix of optimism and fear finds expression in these stories, one presented here. "They will say you cannot, because they could not. They will force you to conform because they did. They will lay down the rules. the ones that they followed. They are everything you can choose not to be because you are more." Some of these stories celebrate the positives of unconventional decisions, while others show the dilemma.

Nostalgia. Memories are precious to any and every person. It has been observed that as the silver haired embrace technology and join social networking sites [7], they seek old friends and catch up. Digital media allow users to store up images and audio clips which they can later revisit and thus relive their memories [13], however the likes and views on such posts can alter how people perceive them. Social media users invariably miss the simple pleasures of the bygone era [16], which finds expression in these narratives, "Lying in the bed, he picks up the same book his mother would. Reads it to our son. I walk in 30 minutes later, find two children, fast asleep."

Automatic clustering themes. Using Top2Vec, three clusters are identified, however, as the presence of similar words have been the basis of this grouping, the stories in each cluster are a mix, and the separation could not be done as desired. In contrast, BERTopic identifies 3 clusters with 75 stories marked as outliers. The first topic clubs those which have some mention of festive celebrations, the second clubs all the stories which mention parents, especially the word mother, and the third is a bag of stories with words like smile and home. In agreement with the results of [9] BERTopic seems to do the clustering better, though the end clusters are still mixed bags.

CONCLUSION

The attempt of clustering 229 miniscule stories with popular techniques Top2Vec and BERTopic highlights the limitations in working with creative texts, and delivers mixed clusters which do align with the themes obtained through thematic analysis: romantic love, self-love, relationships, nostalgia, and choices. Though the field of topic modeling has progressed in leaps and bounds, negative results, such as these, propel further work in natural language processing by machines.

REFERENCES

1. Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 385-393).
2. Ahmed, M. H., Tiun, S., Omar, N., & Sani, N. S. (2022). Short Text Clustering Algorithms, Application and Challenges: A Survey. *Applied Sciences*, *13*(1), 342.
3. Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.
4. Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.
5. Bonsaksen, T., Ruffolo, M., Leung, J., Price, D., Thygesen, H., Schoultz, M., & Geirdal, A. Ø. (2021). Loneliness and its association with social media use during the COVID-19 outbreak. *Social Media+ Society*, *7*(3), 20563051211033821.
6. Buscaldi, D., Schumann, A. K., Qasemizadeh, B., Zargayouna, H., & Charnois, T. (2017, June). Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *International Workshop on Semantic Evaluation (SemEval-2018)* (pp. 679-688).
7. Casanova, G., Abbondanza, S., Rolandi, E., Vaccaro, R., Pettinato, L., Colombo, M., & Guaita, A. (2021). New older users' attitudes toward social networking sites and loneliness: The case of the oldest-old residents in a small Italian city. *Social Media+ Society*, *7*(4), 20563051211052905.

8. Dempsey, A. E., O'Brien, K. D., Tiamiyu, M. F., & Elhai, J. D. (2019). Fear of missing out (FoMO) and rumination mediate relations between social anxiety and problematic Facebook use. *Addictive behaviors reports*, *9*, 100150.
9. Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Frontiers in sociology*, *7*, 886498.
10. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
11. Houlbrook, C., & Armitage, N. (2015). The wishing-tree of Isle Maree: The evolution of a Scottish folkloric practice. *The Materiality of Magic: An Artifactual Investigation into Ritual Practices and Popular Beliefs*, 123-142.
12. Huang, X. (2016, November). Charles Dickens' Critical Realism in David Copperfield. In *4th International Conference on Management Science, Education Technology, Arts, Social Science and Economics 2016* (pp. 1250-1255). Atlantis Press.
13. Jacobsen, B. N., & Beer, D. (2021). Quantified nostalgia: Social media, metrics, and memory. *Social Media+ Society*, *7*(2), 20563051211008822.
14. Korkontzelos, I., Zesch, T., Zanzotto, F. M., & Biemann, C. (2013, June). Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 39-47).
15. Michalopoulos, S., & Xue, M. M. (2021). Folklore. *The Quarterly Journal of Economics*, *136*(4), 1993-2046.
16. Nguyen, P. T. (2017). " Nostalgia for the present": Digital nostalgia and mediated authenticity on Instagram.
17. O'Day, E. B., & Heimberg, R. G. (2021). Social media use, social anxiety, and loneliness: A systematic review. *Computers in Human Behavior Reports*, *3*, 100070.
18. Reddy, M., Methew, M. C., & Kennedy, H. (2021). Social Media: Internet Trends in India and Growth of Social Media in the Recent Times. *Artic Int J Bus Adm Manag Res [Internet]*.
19. Smith, R. (2007, September). An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)* (Vol. 2, pp. 629-633). IEEE.
20. Tierney, G., Bail, C., & Volfovsky, A. (2021). Author Clustering and Topic Estimation for Short Texts. *arXiv e-prints*, arXiv-2106.