Ecological stressors within the River Shannon Catchment

Samuel Dornan

20385853

[samuel.dornan@ucdconnect.ie](mailto:samuel.dornan@ucdconnect.ie)

# **1 Introduction**

Climate change and other human activities have resulted in significant environmental changes across the globe, and Ireland is no exception. In this paper I will attempt to uncover some of the more significant concerns that may be present in Ireland, focusing in particular on the River Shannon Catchment as it is the largest such catchment in Ireland.



*Map of the river Shannon catchment*

The Shannon Catchment has been the subject of significant drainage and development works since the early 20th century, which have had a corresponding effect on the ecology of the region. Specifically, there have been changes in the migratory patterns of local fish such as brown trout, Atlantic salmon, and silver eel among others.

The current ability of most computer programmes to automatically infer significant or extensive information about a topic is limited (Tam et al., 2022, Kandpal et al., 2022) in spite of much fanfare to the contrary (Tshitoyan et al., 2019, Yoshitake et al., 2022), however it is adequate in order to gain a broader understanding of the catchment and any ongoing issues related to it. Specifically, we can use versions of T5 fine-tuned for question answering coupled with targeted questions to gain a broad understanding of the key topics related to the area, including problem areas and major research themes.

To do this, I will use a corpus of text from the Clarivate Web of Science database (Clarivate, 2019), an online database of peer-reviewed articles and papers from a wide variety of publications, which gives users the ability to filter their searches with filters such as the number of citations, region of interest, and publication date among other factors. As well as this, users have the ability to bulk-download article metadata including titles, abstracts, authors, and so forth. This should, in theory, make it possible to gain a better understanding of the text, and by extension the catchment.

**2 Method**

**Preprocessing**

As previously mentioned, in an effort to better understand the Shannon Catchment a corpus of text from the Clarivate Web of Science (Clarivate, 2019) was obtained and cleaned. This cleaning process consisted of removing all entries which had no entries under the Abstract column heading. All other metadata was kept for future reference, though it was not used.

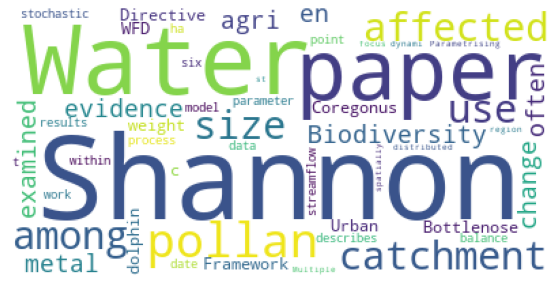
**Analysis**

A version of Google’s T5 Transformer Language Model, fine-tuned for question answering with context (Nunziati, 2022) was used to determine the broad topics covered by each article, with varying degrees of success. These topics were then fed into software to generate a word cloud which would further show the most frequently occurring words in the topics list. (Mueller, 2012)

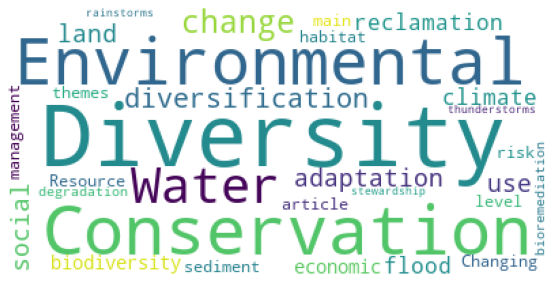
A second word cloud was generated to determine the most frequently occurring themes among the raw abstracts.

# **3 Results**

Key topics of each article were inferred using a version of the T5 transformer model. (Appendix) As well as this, a word cloud was generated for all of the abstracts (Figure 1) and topics (Figure 2) in the dataset.



*Abstracts word cloud*



*Topics word cloud*

The most frequently occurring words in the “Topics” word cloud appear to be “Environmental,” “Diversity,” “Conservation,” and “Water,” while the most commonly occurring words in the “Abstracts” word cloud appear to be “Water,” “Shannon,” and “paper.”

The words “Shannon,” “paper,” and “Water” appear frequently among the abstracts, which may be a result of the dataset being composed of papers about the River Shannon Catchment, as the fact that many abstracts refer to themselves and their focus of study in some way. This could be rectified in the future by manipulating the stopwords list to include words that describe common and somewhat redundant descriptive characteristics, such as “Shannon,” “Water,” and “Paper.”

It does not appear that this change to the stopwords list would have as significant an impact when compared to the abstracts list, given that this was generated using a large language model (LLM).

When used for topic modelling, the LLM appears to perform satisfactorily, determining the broad topics of each of the papers in question. This requires some fine-tuning however, insofar as differing prompts generate different results. As well as this, the specific model must be chosen carefully as differing models are built with differing purposes and use-cases in mind.

Fortunately, tools and model repositories such as HuggingFace (Wolf et al., 2019) contain a large number of pretrained models which can be used off-the-shelf with little or no fine-tuning, as in this case.

# **4 Discussion**

Some of the physically smaller words in the Abstracts word cloud include the words "WFD," and "metal," while some of the physically smaller words in the Topics word cloud include "bioremediation," "risk," and "degradation." These may be interesting topics for further study, if required.

As well as this, some of the content generated by the LLM include the somewhat eye-catching topics of "intersex fish," which upon further research is caused as a result of poorly treated wastewater discharge into the river (McGee et al., 2012), along with a wide variety of GIS methods that may be useful for further research (Appendix).

# **5 Conclusions**

Large language models can be used to aid literature reviews in exceptional circumstances. Automatically generated word clouds can be used to gain further understanding of major themes within a large corpus of literature.

Word clouds, LLMs, and other similar tools can be used to quickly gain a broad overview of the key topics within a large corpus of text, as well as determine areas of future study, however their overall utility appears to be limited to surface-level understanding.

**References**

CLARIVATE, A. 2019. Web of science. *Clarivate Analytics*.

KANDPAL, N., DENG, H., ROBERTS, A., WALLACE, E. & RAFFEL, C. 2022. Large Language Models Struggle to Learn Long-Tail Knowledge. *arXiv preprint arXiv:2211.08411*.

MCGEE, C., BROUGHAM, C., ROCHE, J. & FOGARTY, A. First report of intersex roach residing in Irish rivers downstream of several wastewater treatment plants. Biology and Environment: Proceedings of the Royal Irish Academy, 2012. Royal Irish Academy, 69-77.

MUELLER, A. 2012. *Amueller/word\_cloud: A little word cloud generator in python* [Online]. Available: <https://github.com/amueller/word_cloud> [Accessed 6 December 2022].

NUNZIATI, G. 2022. *Nunziati/bert-vs-T5-for-question-answering: Huggingface-based implementation of an open question answering model trained on the newsqa dataset.* [Online]. Available: <https://github.com/nunziati/bert-vs-t5-for-question-answering> [Accessed 20 November 2022].

TAM, D., MASCARENHAS, A., ZHANG, S., KWAN, S., BANSAL, M. & RAFFEL, C. 2022. Evaluating the Factual Consistency of Large Language Models Through Summarization. *arXiv preprint arXiv:2211.08412*.

TSHITOYAN, V., DAGDELEN, J., WESTON, L., DUNN, A., RONG, Z., KONONOVA, O., PERSSON, K. A., CEDER, G. & JAIN, A. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature,* 571**,** 95-98.

WOLF, T., DEBUT, L., SANH, V., CHAUMOND, J., DELANGUE, C., MOI, A., CISTAC, P., RAULT, T., LOUF, R. & FUNTOWICZ, M. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

YOSHITAKE, M., SATO, F., KAWANO, H. & TERAOKA, H. 2022. MaterialBERT for natural language processing of materials science texts. *Science and Technology of Advanced Materials: Methods,* 2**,** 372-380.

****