Leveraging Machine Learning to Understand Pet Owner Experiences of Feline Pruritus through Social Media Listening

Cherry G[1,3], Mpantis A[2], Rai T[1,3,4], Wright A[5], Wells K [1,3,4]

[1]VHIVE, School of Veterinary Medicine, University of Surrey, Guildford, SRY, UK; [2]ATHENS TECHNOLOGY CENTER (ATC), Athens, Greece; [3]SURREY DATAHUB, School of Veterinary Medicine, University of Surrey, Guildford, SRY, UK; [4]CENTRE FOR VISION, SPEECH AND SIGNAL PROCESSING, University of Surrey, Guildford, SRY, UK; [5]ZOETIS OUTCOMES RESEARCH, Loughlinstown, Co. Dublin, Ireland

INTRODUCTION

Feline pruritus is a common disease in the domestic cat with easily observable symptoms yet its impact on pet owners' lives and quality of life for affected animals remains poorly understood [1]. Traditional research methods, including surveys and interviews, are resource-intensive and necessitate access to representative cohorts, limiting their feasibility. This study harnessed Social Media Listening (SML) to collect pertinent conversations on social media sites. Such data is freely available, public domain and without question bias.

OBJECTIVES

1. Leverage SML data to increase knowledge of feline pruritus and its impact on pet owners and affected cats.
2. Develop a machine learning-driven relevance detection system to automate the identification of relevant social media posts thereby reducing manual labour and human error.
3. Scale data analysis and processing capabilities using advanced natural language processing techniques and fine-tuned models to efficiently analyse large volumes of social media data related to feline pruritus.

METHODS

Keywords, content sources and topics were selected by clinical veterinary dermatology experts and augmented by research literature. Data was collected using ATC's social intelligence platform: Social Asset. Extracting high quality data using SML required well defined relevance criteria with posts manually labelled relevant or irrelevant to create the "ground-truth" input for machine learning.

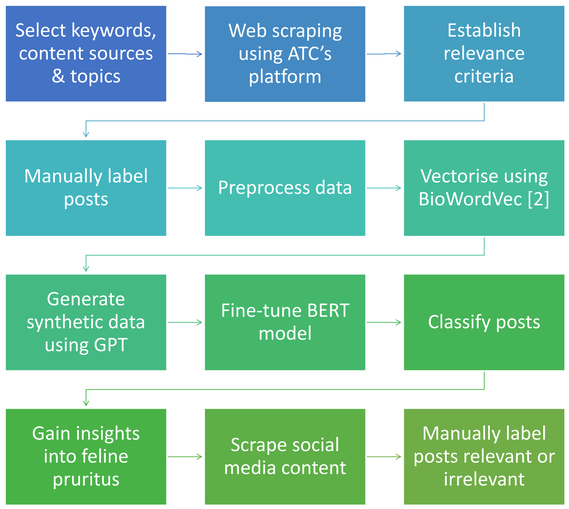

Figure 1 Overall process flow diagram

A dataset comprising 5,000 labelled real-world posts was used from different social media platforms. This was supplemented with 3,800 synthetic posts to mitigate data scarcity for AI training in analysis automation. Data was split into 7,000 training and 1,800 test posts.

Synthetic data was generated using language models like OpenAI GPT. The data captures the style of social media posts. Labelled data considered unsuitable for training due length of post, either too long or too short, were used as input for the LLM.

Data cleaning was applied to Twitter and Reddit posts (real and synthetic) to remove posts of less than seven words before lemmatization and spelling correction. Thorough preprocessing of posts was needed to remove noise.

Posts >500 words were summarised using a GPT language model. BioWordVec model transformed lexicons and incoming documents into vectors [2]. Entities (synonyms and similar words) were extracted using cosine similarity.

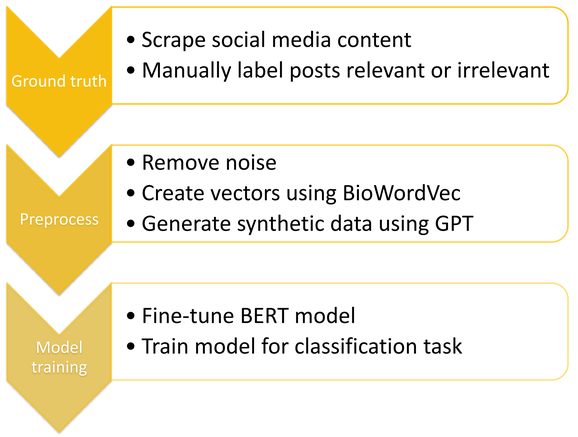

Figure 2 Data preprocessing and model training

RESULTS

A fine-tuned variant of the BERT uncased model [3] for relevance detection, trained on case-specific data over ten epochs, yielded the performance metrics shown in Figure 3.

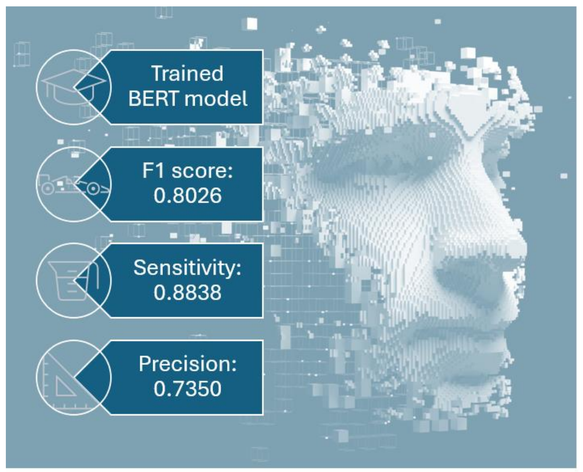

Figure 3 Model performance metrics

The model underwent iterative retraining and revalidation whenever >500 new social media posts were manually labelled. For each iteration, the updated model was compared to the previous one and the best-performing model was retained. Regular evaluation is essential for maintaining effectiveness in a production system.

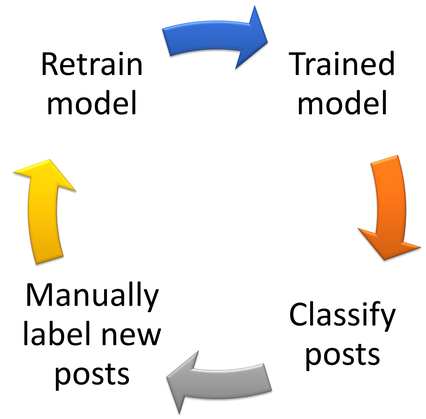

Figure 4 Iterative model validation

CONCLUSIONS

This study highlights the benefits of employing machine learning in relevance detection, reducing human error and marker fatigue while accelerating data analysis scalability. This innovative approach, on manually labelled posts, offers promising insights into feline pruritus and its ramifications on feline and pet owner well-being, potentially validating health-related quality of life measures.

REFERENCES

[1] Williams, S., Cherry, G., Wright, A. et al. (2023). Exploring Symptoms, Causes and Treatments of Feline Pruritus Using Thematic Analysis of Pet Owner Social Media Posts. ICVBD 2023: XVII. International Conference on Veterinary Big Data, 16-17 March, London, UK.

[2] Zhang, Y., Chen, Q., Yang, Z et al. (2019). BioWordVec, Improving Biomedical Word Embeddings With Subword Information and MeSH. *Scientific Data,* 6(1), pp.1–9.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. http://arxiv.org/abs/1810.04805

ACKNOWLEDGEMENTS

This study was wholly funded by Zoetis.