C02029: Doctor of Philosophy
Subject Code: 32903
September 2022

CRICOS Code: 009469A

Computational Understanding of Figurative Language on Social Media

Rhys Biddle

School of Computer Science
Faculty of Eng. & IT
University of Technology Sydney
NSW - 2007, Australia

Computational Understanding of
Figurative Language on Social Media

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Analytics

by

Rhys Biddle

to

Advanced Analytics Institute (AAi)
School of Computer Science
University of Technology Sydney
NSW - 2007, Australia

September 2022

Certificate of Original Authorship

I, *Rhys Biddle* declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

[Rhys Biddle]

Date: 7th September, 2022

Place: Melbourne, Australia

Abstract

Figurative language in online user-generated text poses challenges to Natural Language Processing (NLP) systems designed to automate the understanding of natural language. This thesis introduces empirical studies that quantify the presence and describes the nature of figurative language in Social Media posts (i.e. Twitter). It also quantifies the impact of figurative language on particular NLP applications and introduces new resources (i.e. datasets and methodologies) for the computational processing of figurative language.

This thesis contains a focused case-study on general figurative language in the context of Public Health Surveillance (PHS) applications that monitor Twitter for health events. Findings indicate that some symptom and disease topics are mentioned in a figurative context more than in a health context, which results in a biased signal. To address this bias, a new annotated dataset and text classifier is proposed that reduces bias by targeting figurative expressions of health-related concepts on Twitter.

There is limited research on the expression of hyperbole on Twitter compared to other types of figurative language (e.g., metaphor). To address this gap, a dataset of tweets annotated for the presence of hyperbole is collected and explored. Findings show that hyperbole is relatively common on Twitter and the expression of hyperbole varies from simple and repetitive to complex and novel. A common theme of hyperbole expression on Twitter is the strong affective-laden intentions of the authors, heightening the importance of hyperbole understanding for affective computing applications. Several text classifiers are proposed that leverage pre-trained language models, affective signals, and privileged information for the detection of hyperbole. Experiments show improvements in the detection of hyperbole and importantly highlight annotation biases inherent in the current annotation scheme for hyperbole detection, which is likely to be a roadblock to further improvements.

This thesis quantifies the occurrence of figurative language on Twitter and demonstrates a considerable and consistent presence. Additionally, figurative language is often mishandled by various NLP resources and is scantly addressed by existing datasets and methodologies. Experiment results show that through direct targeting and careful handling of figurative language, improvements to the detection of figurative language are achievable. However, it is concluded that the complexity and novelty of figurative language requires further algorithmic and data inventions for continued progress.

Acknowledgments

This thesis could not have been completed without the wisdom and guidance of many individuals from whom I sought advice throughout the journey. I acknowledge those who had a strong and direct influence on my thinking process and this thesis; Prof. Guandong Xu, Dr. Shaowu Liu, Dr. Aditya Joshi, Dr. Maciek Rybinksi and Dr. Cecile Paris. I also acknowledge my colleagues that helped me better formulate and articulate my ideas in our many conversations over this journey. To the anonymous reviewers that also provided insightful feedback to my submissions throughout my candidature I also extend my acknowledgements to you.

List of Publications

Related to the Thesis :

- **Biddle, R.**, Joshi, A., Liu, S., Paris, C. and Xu, G., 2020, April. *Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter*. In Proceedings of The Web Conference 2020 (pp. 1217-1227).
- **Biddle, R.**, Rybinski, M., Li, Q., Paris, C. and Xu, G., 2021, December. *Harnessing Privileged Information for Hyperbole Detection*. In Proceedings of Australasian Language Technology Association (ALTA).

Others :

- Islam, M.R., Liu, S., **Biddle, R.**, Razzak, I., Wang, X., Tilocca, P. and Xu, G., 2021. *Discovering dynamic adverse behavior of policyholders in the life insurance industry*. Technological Forecasting and Social Change, 163, p.120486.

NOMENCLATURE

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

BiLSTM Bi-Directional Long Short Term Memory Network

BNC British National Corpus

CNN Convolutional Neural Network

ECF Extreme Case Formulation

ELMo Embeddings from Language Model

GloVe Global Vectors for Word Representation

GRNN Gated Recurrent Neural Network

HMC Health Mention Classification

ICD International Statistical Classification of Diseases and Related Health Problems

ICF International Classification of Functioning, Disability and Health

IR Information Retrieval

KNN K-Nearest Neighbour

LDA Latent Dirichlet Allocation

LIWC Linguistic Inquirer Word Count

LSTM Long Short Term Memory Network

MELC Metaphor in end-of-life Care

MIP Metaphor Identification Procedure

NLG Natural Language Generation

NLP Natural Language Processing

NLU Natural Language Understanding

PHS Public Health Surveillance

RNN Recurrent Neural Network

RTT Round Trip Translation

SVM Support Vector Machine

TER Translation Edit Ratio

ULMFit Universal Language Model Fine Tuning for Text Classification

VAD Valence Arousal Dominance

w2v Word2Vec

WHO World Health Organization

Table of Contents

LIST OF FIGURES

List of Tables