

LB/DOAS/137/2019
CS/02/122

EVENT CLASSIFICATION FROM TEXT BASED COMMENTARIES FOR SPORTS ANALYTICS

Kuruppuge Piumi Kanchana Nanayakkara

(158227L)

LIBRARY
UNIVERSITY OF MORATUWA, SRI LANKA
MORATUWA

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science


Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

004¹⁹
004(043)

UNIVERSITY OF MORATUWA
LIBRARY
February 2019

TH 4015
+
CD-ROM

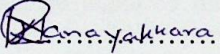
University of Moratuwa

TH4015

TH 4015

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

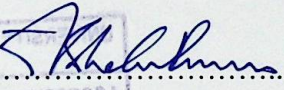
Also, I hereby grant the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or any other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: .....

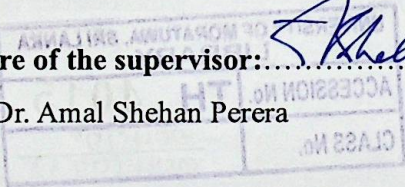
Date:27/05/2019.

Name: K.P.K. Nanayakkara

The above candidate has carried out research for the Masters/MPhil/Ph.D. thesis/dissertation under my supervision.

Signature of the supervisor: ..... Date:27/05/19.....

Name: Dr. Amal Shehan Perera



Abstract

Despite being relatively a new field, which is still evolving, sports analytics have already proven to provide a competitive advantage to sports teams. Especially with large number of commercial leagues taking place across the globe for every sport, the insights generated through analytics has helped franchise owners in bidding auctions for selecting players, where every decision the owners make, will cost them billions of money. However, most of the research work done in sports analytics such as score predictions, player profile analysis has been involved with mainly number crunching, with score boards being the predominant information source of such analytical work.

Over time research had been focused on exploring other sources of information such as video and audio analysis, social media text analysis, micro blogging analysis etc. But there still exists large amount of untapped data with potential of being used for sports analytics. One such data source is online sports commentaries where the web sites give live updates on games, in the form of a temporal series such as minute by minute commentary for soccer or ball by ball commentary for Cricket. These publicly available commentaries give not only the scores, but coverage of all events happening during the match, including the ones which do not go into the scoreboards such as "dropping a catch" in a cricket match. If this information can be captured through event extraction and stored as structured data, that could be useful for analytical purposes.

In this research an end to end system is proposed to produce structured data from online sports commentaries, while extracting information from the commentary text during the process.

Keywords: Sports Analytics, Online Sports Commentaries, Text Mining, Event detection, Multi-label Classification

Acknowledgements

I would like to express my profound gratitude to my supervisor, Dr. Amal Shehan Perera, for introducing this research idea to me and for his invaluable support, encouragement, supervision and useful suggestions throughout this research work. His continuous guidance enabled me to complete my work successfully.

I am as ever, indebted to my family and friends for their continued support throughout my work.

TABLE OF CONTENTS

DECLARATION	I
ABSTRACT	II
ACKNOWLEDGEMENTS	II
TABLE OF CONTENTS	IV
LIST OF FIGURES	VI
LIST OF TABLES	VII
LIST OF ABBREVIATIONS	VIII
CHAPTER 1: INTRODUCTION	1
1.1 BACKGROUND	1
1.2 PROBLEM STATEMENT	2
1.3 RESEARCH OBJECTIVE	4
1.4 SCOPE	4
1.5 CHALLENGES	4
1.6 OVERVIEW OF REST OF THE CHAPTERS	5
CHAPTER 2: LITERATURE REVIEW	6
2.1 OVERVIEW OF TEXT MINING IN SPORTS ANALYTICS	7
2.2 EXISTING WORK ON SPORTS COMMENTARY	8
2.3 EVENT EXTRACTION FROM TEXT	13
2.3.1 Approaches used for event extraction in general	15
2.3.2 Event extraction using multi label classification	19
2.4 ATTRIBUTION OF EVENTS	22
2.5 SUMMARY OF EXISTING WORK	23
CHAPTER 3: METHODOLOGY	24
3.1 OVERALL ARCHITECTURE	24
3.2 IMPLEMENTATION	26
3.2.1 Data Collection	26
3.2.2 Actor Recognition	31
3.2.3 Intermediate Form	31

3.2.4 Dataset Preparation.....	32
3.2.5 Event Classification.....	37
3.2.6 Application Building.....	43
3.3 ADAPTABILITY.....	48
3.4 SUMMARY	49
CHAPTER 4: RESULTS AND ANALYSIS	50
4.1 EXPERIMENT SETUP	50
4.1.1 Dataset	50
4.1.2 Classifier setups	51
4.2 EVALUATION MATRICES	52
4.3 RESULTS AND ANALYSIS	53
4.4 SUMMARY	63
CHAPTER 5: CONCLUSION AND FUTURE WORK	64
5.1 SUMMARY	64
5.2 FUTURE WORK	65
5.3 CONCLUSION.....	65
REFERENCES.....	67

LIST OF FIGURES

Figure 1 Events mentioned in commentary for Cricket.....	3
Figure 2 Algorithm used in [11] for information extraction	9
Figure 3 Algorithm used for web scraping using html parser in [5]	12
Figure 4 Feature vector for a tweet in [3]	18
Figure 5 Key words and phrases used for rule set of [9]	19
Figure 6 Overall Architecture of the system	25
Figure 7 Comparison between Cricbuzz and Cricinfo.....	27
Figure 8 Structure of json object.....	30
Figure 9 No of commentary sentences per event.....	40
Figure 10 ER diagram	48
Figure 11 Comparison of F1 Score vs No. of labels in training set	58
Figure 12 Results for OnevsRest Linear SVM in [4].....	60
Figure 13 Label distribution of [4].....	61
Figure 14 Accuracy of the classifiers tested in [5].....	62

LIST OF TABLES

Table 1 Key Word dictionary used to validate manual tags	37
Table 2 Multiple tags per Commentary sentence	39
Table 3 Roles associated with all events	44
Table 4 Table Design.....	47
Table 5 Label counts in Dataset	51
Table 6 Average performance of all set ups across classes	54
Table 7 Performance of Multinomial NB with rare categories	55
Table 9 F1 scores per Label	56
Table 10 Performance of algorithms in different problem transformation methods..	61
Table 11 Accuracy scores of best performing classifier set ups.....	62

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BR	Binary Relevance
BOW	Bag of Words
BI	Business Intelligence
CC	Classifier Chains
ECC	Ensembles of Classifier Chains
EPS	Ensembles of Pruned Sets
GNB	Gaussian Naive Bayes
IE	Information Extraction
LP	Label Power Set
LBW	Leg Before Wicket
ML	Machine Learning
NB	Naive Bayes
NER	Named Entity Recognition
ODI	One Day International
POS	Part of Speech
RNN	Recurrent Neural Network
RE	Regular Expressions
SVC	Support Vector Classifier
SVM	Support Vector Machine
TF-IDF	Term Frequency – Inverse Document Frequency