





Dhinaharan Nagamalai  
Brajesh Kumar Kaushik (Eds)

# Computer Science & Information Technology

4<sup>th</sup> International Conference on Natural Language Computing (NATL 2018)  
April 28~29, 2018, Dubai, UAE.



**AIRCC Publishing Corporation**

## **Volume Editors**

Dhinaharan Nagamalai,  
Wireilla Net Solutions, Australia  
E-mail: dhinthia@yahoo.com

Brajesh Kumar Kaushik,  
IIT-Roorkee, India  
E-mail: bkkaushik23@gmail.com

ISSN: 2231 - 5403  
ISBN: 978-1-921987-84-7  
DOI : 10.5121/csit.2018.80601 - 10.5121/csit.2018.80612

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

## Preface

The 4<sup>th</sup> International Conference on Natural Language Computing (NATL 2018) was held in Dubai, UAE during April 28~29, 2018. The 4<sup>th</sup> International Conference on Computer Science, Engineering and Applications (CSEA 2018), The 4<sup>th</sup> International Conference on Data Mining and Database Management Systems (DMDBS 2018), The 4<sup>th</sup> International Conference on Fuzzy Logic Systems (Fuzzy 2018), The 4<sup>th</sup> International Conference on Information Technology Converge Services (ITCON 2018), The 2<sup>nd</sup> International Conference on Networks and Security (NSEC 2018) and The 2<sup>nd</sup> International Conference on Computer Science and Information Technology (COMIT 2018) was collocated with The 4<sup>th</sup> International Conference on Natural Language Computing (NATL 2018). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The NATL-2018, CSEA-2018, DMDBS-2018, Fuzzy-2018, ITCON-2018, NSEC-2018, COMIT-2018 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, NATL-2018, CSEA-2018, DMDBS-2018, Fuzzy-2018, ITCON-2018, NSEC-2018, COMIT-2018 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the NATL-2018, CSEA-2018, DMDBS-2018, Fuzzy-2018, ITCON-2018, NSEC-2018, COMIT-2018.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Dhinaharan Nagamalai  
Brajesh Kumar Kaushik

## Organization

### General Chair

David C. Wyld  
Jan Zizka

Southeastern Louisiana University, USA  
Mendel University in Brno, Czech Republic

### Program Committee Members

Abdulbaset Mohammad	University of Bradford, United Kingdom
Ahmed M. Khedr	Sharjah University, UAE
Ahmed Nada	Al-Quds University, Palestine
Alejandro Garces	Jaume I University, Spain
Amin Zadeh Shirazi	Islamic Azad University of Mashhad, Iran
Amine Achouri	University of Tunis, Tunisia
Andino Maselena	STMIK Pringsewu, Indonesia
Anja Richert	RWTH Aachen University, Germany
Ankit Chaudhary	Truman State University, USA
Atif Farid Mohammad	University of North Carolina, Charlotte
Ayad Salhie	Australian College of Kuwait, Kuwait
Baghdad ATMANI	University of Oran, Algeria
Bhuyan Jay	Tuskegee University, United States
Bo-Hao Chen	Yuan Ze University, Taiwan
Buket Barkana	University of Bridgeport, USA
Cherif Foudil	LESIA Laboratory Biskra University, Algeria
Chin-Chih Chang	Chung Hua University, Taiwan
Chiranjib Sur	University of Florida, US
Chris Panagiotakopoulos	University of Patras, Greece
Diego Reforgiato	University of Catania, Italy
Dongchen Li	Peking University, China
Dongpo Xu	Northeast Normal University, China
Emilio Jimenez Macias	University of La Rioja, Spain
Ermias Birihanu	Wolkite University, Ethiopia
Farzad Kiani	Istanbul S.Zaim University, Turkey
Fatih Korkmaz	Cankiri Karatekin University, Turkey
Fazlul Haque K.M	Daffodil International University, Bangladesh
Fernando Zacarias	Universidad Autonoma de Puebla, Mexico
Ferran Torrent	Universitat de Girona, Girona
Ghazi Al-Naymat	University of Dammam, Saudi Arabia
Haibo Yi	Shenzhen Polytechnic, China
Hamza Aldabbas	De Montfort University, United Kingdom
Hao-En Chueh	Yuanpei University of Medical Technology, Taiwan
Harmas Mohamed	University of Setif, Algeria
Hemant Kumar Reddy K	National Institute of Science and Technology, India
Himanshu Mehta	Telecom Paris Tech, India
Hoshang Kolivand	Universiti Teknologi Malaysia, Malaysia

Hossein Jadidoleslamy	MUT University, Iran
Intisar Al-Mejibli	University of Essex, United Kingdom
Isa Maleki	Islamic Azad University, Iran
ISAMM	University of Manouba, Tunisia
Izzat Alsmadi	Damascus University, Syria
Jacques Demerjian	Communications & Systems, France
Jerin Cyriac ME	Truman State University, USA
Jingyan Wang	New York University, UAE
Jose Vicente Berna Martinez	University of Alicante, Spain
Ka Chan	La Trobe University, Australia
Laiali Almazaydeh	University of Bridgeport, USA
Marina Marjanovic Jakovljevic	Singidunum University, Serbia
Martins Irhebhude	Nigerian Defence Academy, Nigeria
Maryam Rastgarpour	Islamic Azad University, Iran
Md. Shahjahan Ali	Islamic University, Bangladesh
Melih Kirlidog	Marmara University, Turkey
Mellal Mohamed Arezki	M'Hamed Bougara University, Algeria
Ming Fan	Broadcom Corporation, USA
Mohamad heidari	Islamic Azad University, Iran
Mohammad Jafarabad	Qom University, Iran
Mohammad Talib	University of Botswana, Botswana
Murat Canayaz	Yuzuncu Yil University, Turkey
Nidal M. Turab	Al-Isra University, Jordan
Nikita Barabanov	North Dakota State University, Fargo
Nishant Doshi	PDPU Gandhinagar, India
Ognjen Kuljaca	Alcorn State University, USA
Oluwatobi Olabiyi	Digital Compression Technology, USA
Poo Kuan Hoong	Multimedia University, Malaysia
Prof. Bimal K. Bose	University of Tennessee, USA
Robert Burduk	Wroclaw University of Technology, Poland
Saad M. Darwish	Alexandria University, Egypt
Salem Nasri	Monastir University, Tunisia
Sameh Abd EL-Haleem	Menoufia University, Egypt
Samy Abu Naser	Al Azhar University, Gaza, Palestine
Shengqian Yang	Ohio State University, USA
Sonali Vyas	Amity University Rajasthan, India
Souad Zid	National Engineering School of Tunis, Tunisia
Stephan	Alpen-Adria Universitat Klagenfurt, Austria
Tad Gonsalves	Sophia University, Japan
Thandar Thein	University of Computer Studies, Myanmar
Tien D. Nguyen	Coventry University, United Kingdom
Vidroha Debroy	Hudson Alley Software, USA
Vikram Puri	DuyTan University, Vietnam
Wembe Sop Diake Hubert	University of Douala, Cameroon
Wernhuar Tarnng	National Hsinchu University of Education, Taiwan
Woo Chaw Seng	University of Malaya, Malaysia
Wtarnng	National Hsinchu University, Taiwan
Yacef Fouad	Division Productique et Robotique, Algeria
Zamira Daw	United Technologies Research Center in Berkeley, USA

## **Technically Sponsored by**

**Computer Science & Information Technology Community (CSITC)**



**Networks & Communications Community (NCC)**



**Soft Computing Community (SCC)**



## **Organized By**



**Academy & Industry Research Collaboration Center (AIRCC)**



## TABLE OF CONTENTS

### 4<sup>th</sup> International Conference on Natural Language Computing (NATL 2018)

**Neural Symbolic Arabic Paraphrasing with Automatic Evaluation** ..... 01 - 13  
*Fatima Al-Raisi, Abdelwahab Bourai and Weijian Lin*

**Applying Distributional Semantics to Enhance Classifying Emotions in Arabic Tweets** ..... 15 - 34  
*Shahd Alharbi and Matthew Purver*

**General Regression Neural Network Based PoS Tagging for Nepali Text** ..... 35 - 40  
*ArchitYajnik*

**Social Network Hate Speech Detection for Amharic Language**..... 41 - 55  
*Zewdie Mossie and Jenq-Haur Wang*

**Importance of Verb Suffix Mapping in Discourse Translation System**..... 143 - 151  
*Suryakanthi Tangirala*

### 4<sup>th</sup> International Conference on Computer Science, Engineering and Applications (CSEA 2018)

**Social Media Analytics for Sentiment Analysis and Event Detection in Smart Cities** ..... 57 - 64  
*Aysha Al Nuaimi, Aysha Al Shamsi and Amna Al Shamsi, Elarbi Badidi*

**Character and Image Recognition for Data Cataloging in Ecological Research** ..... 65 - 76  
*Shannon Heh*

### 4<sup>th</sup> International Conference on Data Mining and Database Management Systems (DMDBS 2018)

**Probability Based Cluster Expansion Oversampling Technique for Imbalanced Data** ..... 77 - 90  
*Shaukat Ali Shahee and Usha Ananthakumar*

## **4<sup>th</sup> International Conference on Fuzzy Logic Systems (Fuzzy 2018)**

**Validation Method of Fuzzy Association Rules Based on Fuzzy Formal  
Concept Analysis and Structural Equation Model ..... 91 - 108**  
*Imen Mguiris, Hamida Amdouni and Mohamed Mohsen Gammoudi*

## **4<sup>th</sup> International Conference on Information Technology Converge Services (ITCON 2018)**

**Classification of Alzheimer Using fMRI Data and Brain Network..... 109 - 119**  
*Rishi Yadav, Ankit Gautam, Ravi Bhushan Mishra*

## **2<sup>nd</sup> International Conference on Networks and Security (NSEC 2018)**

**Automated Penetration Testing : An Overview..... 121 - 129**  
*Farah Abu-Dabaseh and Esraa Alshammari*

## **2<sup>nd</sup> International Conference on Computer Science and Information Technology (COMIT 2018)**

**Exact Solutions of a Family of Higher-Dimensional Space-Time Fractional  
KDV-Type Equations..... 131 - 141**  
*Mohammed O.AL-AMR*

# NEURAL SYMBOLIC ARABIC PARAPHRASING WITH AUTOMATIC EVALUATION

Fatima Al-Raisi, Abdelwahab Bourai and Weijian Lin

Language Technologies Institute, School of Computer Science,  
Carnegie Mellon University, Pittsburgh, PA, USA

## ABSTRACT

*We present symbolic and neural approaches for Arabic paraphrasing that yield high paraphrasing accuracy. This is the first work on sentence level paraphrase generation for Arabic and the first using neural models to generate paraphrased sentences for Arabic. We present and compare several methods for para-phrasing and obtaining monolingual parallel data. We share a large coverage phrase dictionary for Arabic and contribute a large parallel monolingual corpus that can be used in developing new seq-to-seq models for paraphrasing. This is the first large monolingual corpus of Arabic. We also present first results in Arabic paraphrasing using seq-to-seq neural methods. Additionally, we propose a novel automatic evaluation metric for paraphrasing that correlates highly with human judgement.*

## KEYWORDS

*Natural Language Processing, Paraphrasing, Sequence-to-Sequence Models, Neural Networks, Automatic Evaluation, Evaluation Metric, Data Resource*

## 1. INTRODUCTION

Paraphrasing and paraphrase detection are two important problems in natural language processing. Paraphrasing-based applications include text simplification and text generation from structured knowledge [19]. Other paraphrastic models include machine translation and sentence summarization [14, 24, 6]. Paraphrases are useful not only in generation tasks but also in analysis tasks such as information retrieval and question answering [25, 20, 9].

We present and compare two different approaches for sentence paraphrasing in Arabic: a phrase-based method and a neural method. To our knowledge, this is the first work on sentence paraphrasing for modern standard Arabic.

We also present a novel approach for obtaining parallel monolingual data and use the acquired data to train our neural sequence-to-sequence model. As a by-product of this work, we contribute a large parallel monolingual corpus for Arabic containing two million sentence pairs which can be used to develop new seq-to-seq models for paraphrasing. We also build a phrase database for Arabic containing over 88K phrase pairs of various lengths. Another contribution of our work is

devising and testing a new evaluation metric for paraphrasing. We present encouraging initial results using this metric in this paper. The remainder of this paper is structured as follows: we contextualize our work within paraphrasing research in Section 2, we present the phrase-based and neural approaches for paraphrasing sentences and building phrase dictionaries in sections 3 and 4. We present details and discuss experiments on the evaluation metric in Section 6. We conclude with plans for future extensions of the work.

## 2. RELATED WORK

The paraphrase database project PPDB has paraphrase resources for multiple languages [3], including Arabic. The paraphrases are obtained using parallel bilingual corpora by applying the pivot method where one language is used as a bridge or intermediate meaning representation [3]. Paraphrases from dialectal Arabic to standard Arabic have been used in [21] to improve Arabic-English statistical machine translation. Turker assisted paraphrasing has been used in [8] to improve English-Arabic MT. A comparison between various paraphrase acquisition techniques on sentential paraphrasing is given in [5] but does not include experiments on Arabic sentential paraphrasing.

## 3. EXTRACTING PARAPHRASES FROM BILINGUAL DATA

Our first approach to Arabic paraphrasing is the pivot method proposed by Bannard and Callison-Burch [3]. A key benefit is that it is language-agnostic and is based on the idea that any two source strings  $e_1$  and  $e_2$  that both translate to a reference string  $f_1$  have similar meaning. Bannard and Callison-Burch used English as the reference string  $f$ , but in our study we will instead pivot into English to obtain paraphrase pairs [3]. We obtain the final paraphrase probability by marginalizing over the English translation probabilities with  $e$  and Arabic phrases  $a_1$  and  $a_2$ . A mathematical formulation of the approach can be found in Equation 1.

$$p(a_2|a_1) = \sum_e p(a_2|e)p(e|a_1) \quad (1)$$

In order to extract paraphrases, we first obtained a parallel bilingual corpus through English and Arabic versions of the EUROPARL dataset [13]. We pruned the corpus to only contain sentences with less than 80 words and tokenized using the StanfordNLP Arabic Tokenizer [15]. This was achieved in the data preprocessing step by computing the sentence length and excluding sentences with more than 80 tokens. This gave us a final corpus size of 241,902 sentences.

Additionally, to calculate conditional probabilities for our paraphrase equation, we need alignment. Thus we ran GIZA++ [18], the well-known alignment tool widely used in MT, to obtain these alignments [18]. We chose GIZA++ for its previous success with machine translation involving Arabic [1]. Once we have a database of paraphrase mappings, we can then substitute phrases with their corresponding paraphrases by selecting the phrase with the highest probability. This substitution approach was used by Bannard and Callison-Burch in their study as well [3] The way we extract the paraphrase is summarized in Equation 2. An example of this process can be seen in Figure 1

$$\hat{a}_2 = \operatorname{argmax}_{a_2 \neq a_1} p(a_2|a_1) \quad (2)$$



Figure 1: An example paraphrased sentence produced using the pivot method.

### 3.1 Improving Coverage of Phrase Database

In our initial experiments, we noticed that generated phrase pairs did not necessarily match in some grammatical features such as definiteness and number. We post-processed the phrase dictionary to add entries with other variants of the phenomenon for completion. For example, for a word pair that appears in the phrase table where one word is definite and the other is not, we add two entries where both are definite and both are indefinite. We did not adjust the scores to reflect this but ordered the entries according to observed frequency of the word/phrase. For definiteness, we limited the addition to the clear definite marker “al-” in Arabic. We applied the same for simple cases of number matching where the morphology is concatenative or easily processed. We note that this modifications did not include all possible mismatches since they were based on simple heuristics. However, this may have contributed to better grammaticality as discussed in Section 5. We also noticed cases like the following in the generated phrase table:

x ||| y ||| score

x ||| z ||| score

We included, for improved coverage, the following entry:

y ||| z ||| score

Again, this was limited in scope since we relied on simple string match to identify such entries.

### 3.2 Phrase-substitution

We randomly sampled 100 sentences from the datasets we have [23, 12] and performed phrase substitution. Figure 2 shows a sample of paraphrased sentences acquired using phrase substitution. We note that in the last example the output sentence differs from the original in only one word (last word) but the meaning is entirely altered. We discuss results and experiments on the quality of paraphrased sentences in Section 5.

<b>Original</b>	الاهتمام بوضع المرأة يقفز مجدداً الى الواجهة في السعودية
<b>Paraphrased</b>	الاهتمام بوضع المرأة يقفز مرة أخرى في المقدمة في المملكة
<b>Original</b>	وقال النعيمي ان المضاربين على البترول هم اصحاب تأثير مهم على اسعاره .
<b>Paraphrased</b>	وقال النعيمي ان المضاربين على النفط هم اصحاب تأثير مهم على ثمنه .
<b>Original</b>	أعلن الدكتور كينيث أليس رئيس هيئة المعونة الأمريكية، توقف الولايات المتحدة قريباً عن تمويل مشروعات البنية الأساسية في مصر
<b>Paraphrased</b>	كشف الدكتور كينيث أليس رئيس لجنة المساعدة الأمريكية، توقف الولايات المتحدة عما قريب عن تمويل مشاريع البنية التحتية في مصر
<b>Original</b>	سيتضمن الوفد عدداً من الشركات المعنية بالإنتاج والتصدير الزراعي للتعرف على الآليات المتبعة في هيئة الرقابة على الصادرات الزراعية الطازجة بجنوب أفريقيا لضمان جودة المنتجات الزراعية واعتمادها دولياً .
<b>Paraphrased</b>	سيتضمن الوفد عدداً من الشركات ذات الصلة بالإنتاج والتصدير الزراعي للتعرف على التدابير المتبعة في هيئة الرقابة على الصادرات الزراعية الطازجة بجنوب أفريقيا لضمان نوعية المنتجات الزراعية واعتمادها دولياً .
<b>Original</b>	6.1 بليون دولار إجمالي الديون . المصارف المصرية ترفض مقايضة ديون المتعثرين بمشاريعهم العقارية
<b>Paraphrased</b>	6.1 بليون دولار إجمالي الديون . المصارف المصرية ترفض مقايضة ديون المتعثرين بمشاريعهم الدوائية

Figure 2: Sample output produced using phrase substitution

#### 4. MONOLINGUAL PARALLEL DATA FOR SEQ-TO-SEQ PARAPHRASING

We need parallel monolingual data to train our sequence-to-sequence paraphrasing model. To address the lack of parallel monolingual data for Arabic, we propose a novel method for generating Arabic parallel language sentences using two other language pairs as resource. The idea is to use paired sentences data from two other languages, translate them into Arabic correspondingly, then use them as the training data for sequence-to-sequence machine translation model to train a translator to generate Arabic to Arabic paraphrases.

The first advantage of this approach is its scalability. After preparing enough training data for the paraphrase model, the generation step for Arabic paraphrases is easily scalable. The second advantage is that the seq-to-seq paraphraser model may contain valuable insight for building Arabic paraphrases database at words and phrases level, since state of the art neural machine translation techniques are capable of capturing word and phrase level similarity by projecting word embeddings into vector space.

We used europarl-v7 fr-en [13] data which contains two million sentence pairs, and then used Google translate API to generate French-Arabic and English-Arabic sentence pairs

correspondingly. Then we paired the output to construct parallel monolingual data and used it as training data for a Bi-LSTM with embedding size and hidden size set to be 512 and attention size set to be 128. The bidirectional LSMT model is chosen to create a stronger representation that takes into account patterns found in the right-to-left Arabic sentence direction as well as features and patterns present towards the end of the sentence which are attenuated in right-to-left unidirectional representation.

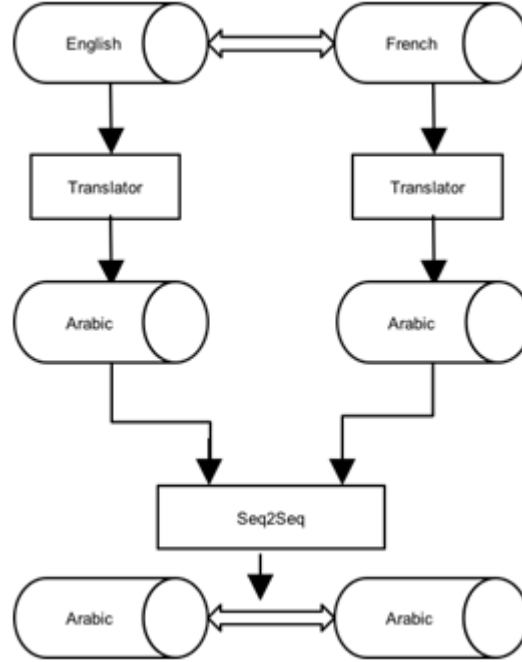


Figure 3: Overview of the process completed to obtain two parallel Arabic corpora generated from different source languages

The grammaticality of the Arabic output sentences were independently evaluated by two speakers of Arabic and found to be both grammatical and semantically coherent. The process is demonstrated in Figure 3. Training the Bi-LSTM took about 7 days on this dataset and we obtained a corpus of monolingual Arabic containing two million parallel sentences.

## 5. RESULTS

We report results on the bilingual pivot and sequence-to-sequence approaches detailed above.

### 5.1 Phrase-based Method

Using the pivot method, we obtained over 88K phrase pairs. We report a few results. First, Figure 4 shows the length distribution for phrases in the database.

We obtained human evaluations from two native speakers on the grammaticality and meaning preservation aspects of the paraphrased sentences using a sample of 200 sentence pairs. We note that this sample is more than two times larger than samples used in similar evaluations in

related work [17]. For each criterion, we asked the annotator to judge the quality of the output on a scale from 1 to 5. We chose this scale to capture variation in the level of grammaticality (since there are minor and more serious grammatical mistakes) and in the extent to which the paraphrased sentence preserved the meaning of the original sentence. The agreement between annotators calculated in terms of IntraClass Correlation, preferred for ordinal data, is summarized in Table 1. It was not expected to find higher agreement on meaning preservation since it is more subjective than grammaticality. It is possible that phrases substituted were of similar meaning yet possibly resulted in unusual sentence structure which made the meaning preservation judgement straightforward while grammaticality harder to judge. We analyzed changes made to reference sentences that received the highest score in meaning preservation after paraphrase substitution (score = 5) . A change is a word replacement or deletion. Table 2 summarizes these changes. When  $n$  consecutive words are replaced by  $n$  or more words, we consider those to be  $n$  changes rather than 1 phrasal change. Table 3 summarizes the evaluation of paraphrased sentences obtained using phrase substitution.

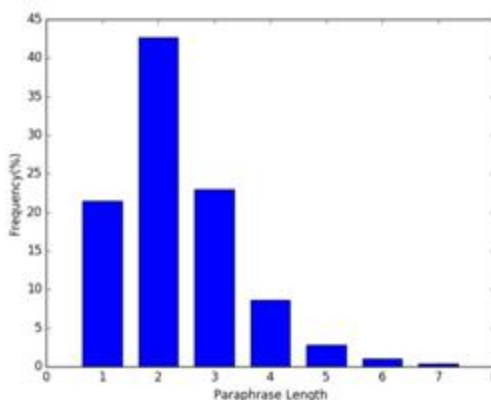


Figure 4: Distribution of phrase lengths in our paraphrases obtained through pivoting

Table 1: Inter-annotator agreement

	ICC (single)	ICC (average)
Grammaticality	0.375	0.545
Meaning	0.707	0.547

## 5.2 Neural Method

The corpus of sentence pairs produced was large (2M). By the time the output was produced, our human evaluation resources were exhausted, so we qualitatively evaluate it on a subset consisting of 200 sentence pairs. The monolingual parallel sentences obtained were found to be grammatical or near grammatical with minor mistakes but were found to be highly similar in meaning. Also, they were sufficiently different in surface form which encouraged their use as training data for the a seq-to-seq paraphrasing model. Initially, we were concerned of the possibility that the en→ar and fr→ar translation tools we used were trained on the same Arabic data with similar model architecture and parameters, in which case the output would not be diverse enough to be useful as input to the paraphrasing system. We were also concerned of the possibility that the French source was translated to English and then to Arabic. However, judging by the diversity in the output of these models, using the resultant parallel monolingual corpus to



train the neural paraphrasing model is justified. We compare the surface diversity in the output of the fr-ar and en-ar MT models by computing word and phrase overlap and find enough surface variation to train the model adequately.

Table 2: Number of changes in paraphrased sentences with high meaning preservation score

No.of Changes	5	4	3	2	1
Frequency (Sent)	3	4	5	16	26

Table 3: Evaluation of paraphrased sentences (averaged and rounded from two annotator ratings)

Scale	Grammaticality %	Meaning %	Both %
5	62	62	50
4	21	16	7
3	10	9	3
2	7	12	4
1	0	1	0

<b>Original</b>	بولين توفيق حصول شركة اميركية على رخصة تصنيع دبابه "ليوبارد" الالمانية
<b>Paraphrased</b>	وعلاوة على ذلك، فإن المحكمة الجنائية ترفض حصول حصول توفيق رخصة الالمانية
<b>Original</b>	1945 - افتتاح مؤتمر بوتسدام بين قادة الاتحاد السوفياتي وبريطانيا والولايات المتحدة الذي رسم حدود الدول الاوروبية بعد الحرب العالمية الثانية
<b>Paraphrased</b>	وعلاوة على ذلك، يبدو أن افتتاح بين الاتحاد الأوروبي والولايات المتحدة والاتحاد الأوروبي والولايات المتحدة في الاتحاد الأوروبي والولايات المتحدة التي من شأنها أن تؤدي إلى منظمة التجارة العالمية العالمية
<b>Original</b>	1971 - حسين الملك حسين يلقي يلقي الاتفاقات الاتفاقات الاتفاقات الاتفاقات
<b>Paraphrased</b>	1971 - العامل الاردني الملك حسين يلقي الاتفاقات التي تسمح لتتلتطيمات الفلسطينية بإقامة قواعد لها في الاردن

Figure 5: Sample output from neural paraphrasing model

The paraphrased output from the neural system was far from grammatical and in some cases the meaning was incomplete due to early sentence truncation, especially for long sentences. Better output was observed for shorter sentences which conforms with the pattern typically seen in neural seq-to-seq models. However, we make the following observations about the output:

- The model learns what phrases to use at the beginning of the sentence. It uses things like "Moreover" and "As you know," (translated from Arabic), exactly at the beginning of the sentence.
- The model seem to learn and include central parts of the sentence in the output such as the subject or the location of the event.
- The model learns correspondences between the main parts of the sentence; e.g., the byline vs. remainder of the sentence and quoted text vs. part before the quotation.
- The model often fails in producing output with the correct word order.
- As observed with neural language models, it tends to repeat words.

Figure 5 shows a sample output from the neural model.

## 6. AN AUTOMATIC EVALUATION METRIC

Since human evaluation is time-consuming, subjective and not always available, we propose an automatic evaluation metric for paraphrasing. We propose criteria for judging paraphrase quality and operationalize those criteria using well-defined functions. A good paraphrase has the following two properties:

1. maintains the meaning of the original text, yet
2. expresses the same meaning using different surface realizations.

To evaluate the semantic similarity and surface variation in a paraphrase we employ well-defined metrics discussed next.

### 6.1 Semantic Similarity

Several methods exist for capturing the semantic similarity of text [11, 7, 2]. One simple approach uses the distributional properties of words in the sentence and embeds a sentence by averaging the embedding vectors of its words. We choose this method for its simplicity and efficiency. The sentence vector is thus given by:

$$w_x = \frac{1}{|x|} \sum_1^n w_{x_i} \quad (3)$$

where  $n$  is the number of words in the sentence. Although this method is simple and does not consider word-order or syntactic structure, it performs surprisingly well when compared to more advanced neural- based methods designed to capture sentential semantics [22]. It also correlates highly with human judgement [22]. Also, being agnostic to word order is actually a desired property in the paraphrase case since valid paraphrases may only differ in the order of the words or the construction used. For example, in Arabic the SVO word order can almost always be changed to VSO without changing the meaning of the sentence (except for emphasis) and without introducing any other particle. Another example is English active voice and the corresponding passive voice sentence ( + by Subj) of the original sentence.

To compute the semantic similarity between two sentences, the original sentence and the paraphrase, the cosine similarity between the two sentence vectors is computed. In our experiments, we use word embeddings with 300 dimensions trained on Arabic text from [4]. Cosine similarity is chosen for its efficient and parallelizable computation. An evaluation metric should be computed efficiently for it to be useful in evaluating output and comparing several systems performance in real-time. Note that the purpose of the metric is not to capture the most accurate distance between the vectors in space per se but to efficiently and reasonably estimate the similarity between the sentences in a principled way which cosine similarity suffices for. In our generic framework, this metric could be replaced with any other semantic similarity metric of choice.

## 6.2 Surface Variation

To capture surface variation we first map each sentence into a common vocabulary space and compute the hamming distance between the sentence vectors in that space. This also limits sentence length bias where short sentences will naturally have less surface overlap. In our experiments, we map sentences into a vocabulary space of 610977 words [4]. We present experiments and results in section 6.4.

## 6.3 Combining Criteria for Meaning and Form

Minimal change in surface form can result in maximal preservation of original meaning. However it will score low on surface variation. Similarly, if surface form is significantly changed, we may risk altering the meaning of the original sentence. Since semantic similarity and surface variation are two competing criteria, we combine them using (balanced) harmonic mean. The final score of the paraphrase is given by:

$$s = 2 \frac{\text{SemanticSimilarity} \cdot \text{LexicalDistance}}{\text{SemanticSimilarity} + \text{LexicalDistance}} \quad (4)$$

## 6.4 Results

We evaluate a set of 198 Arabic sentence pairs sampled from newswire data [12] as follows. These include headlines on similar topics and for each headline one or two sentences detailing the event in the headline or reiterating it. Pairs including the headline and the following sentence were reasonable paraphrase pairs whereas the other pairs varied in paraphrasing potential from moderate (some overlap in meaning) to poor (unrelated, contradicting or little overlap). We created 576 sentence pairs from the dataset but obtained annotations for only 198 of them. This evaluation of the metric was conducted before obtaining paraphrasing results from our phrase-based and neural models. Therefore, we created sentence and paraphrase pairs following this approach. Since sentences were obtained from newswire data, we assumed they are grammatical and did not obtain grammatical judgement from annotators. On paraphrastic quality, human evaluations were obtained from three annotators who are native speakers of Arabic. Each annotator was asked to judge the quality of the paraphrase, on whether it preserved meaning and was expressed differently, on a ordinal scale from 1 to 5 where 1 indicates poor quality. We used R to measure interannotator (absolute) agreement using IntraClass Correlation (ICC) and agreement was measured at 0.714 ICC which is considered “very good.” The biserial correlation between the binarized human evaluations and the evaluation metric scores was 0.813.

## 6.5 Analysis

Observing high correlation between human evaluation and the proposed evaluation metric, we examined the dataset to see if results were biased by sampling or data peculiarities. For sentence pairs including the headline and the following sentence, both human and evaluation metric scores were high. For most of the other sentence pairs, the paraphrase was judged as weak or poor. In both of these cases, the judgement was “easy” and straightforward and this perhaps lead to the surprisingly good results. Perhaps sentence pairs with finer and more subtle semantic phenomena such as polysemy and synonymy would have been harder to score accurately by the metric. We need to conduct more experiments to verify this.

We also explored the effect of the embedding dimension on the correlation between human evaluation and the metric we proposed. We experimented with lower dimensions: 50, 100, 150, 200, 250 and obtained slightly lower values of biserial correlation as we decreased the embedding dimension. With 50 dimensions, the absolute difference between the previous biserial correlation value and the new one was 0.03. It is worth mentioning that when only using overlap as a measure of semantic similarity, the correlation between human judgement and the evaluation metric is 0.47. We clearly gain by using word embeddings but even something as simple as word overlap can capture semantic distance to some extent and explain a good amount of variation in human judgement.

We initially set out to compute the surface distance between two sentences using a measure that only depends on the two sentences as input; such as token/character overlap or minimum edit distance. However, since this measure can have a length bias we use the more robust hamming distance which is computed in a canonical space.

Using a common canonical space for doing the various computations is preferred as these spaces comprise a reference against which candidates are evaluated. When using the metric to evaluate outputs from different systems, the reference can be decided at test time to avoid “gaming” the metric. It is desirable to have a metric that does not require a reference sentence when evaluating a candidate paraphrase since 1. results and rankings can be sensitive to the choice of the reference sentence which is often subjective and 2. the notion of a “reference paraphrase” is problematic here since paraphrasing is essentially based on divergence from a given surface form while preserving the underlying meaning and hence is loosely constrained. However, we do recognize the problem with not having a reference for an evaluation metric: it can make it susceptible to “gaming” by players competing to optimize the metric score. Therefore, we propose to decide the vocabulary space in which surface distance function is computed and the semantics space in which semantic similarity is computed at test time.

Also, since this metric combines competing objectives, the parameter controlling the relative strength of each component can also be decided at test time to improve the metric robustness. While it has been shown that it is possible to find a Pareto optimal hypothesis that aims to jointly optimize several different objectives [10], we argue that those objectives are not exactly competing or orthogonal. They may be weakly correlated but they are still positively correlated since they compare surface distance against the same reference (as in BLUE and TER), which is not the case in our proposed metric setting.

## 7. CONCLUSION

We presented and compared two different approaches for sentence paraphrasing in Arabic. The phrase-based approach yielded very good results when used to create paraphrase sentences. The neural method output is still far from practical applicability but the model has learned interesting linguistic constructs like phrases used for sentence opening. We also presented a novel approach for obtaining parallel monolingual data and contributed a dataset of two million parallel sentence pairs in Arabic using this approach. We applied the pivoting method to construct a large coverage paraphrase database for Arabic that includes over 88K phrase pairs. When used to create new sentences, the paraphrase dictionary gave very good results on grammaticality and meaning preservation. We proposed a new automatic evaluation metric for paraphrasing that

does not require the use of a reference sentence while evaluating candidate hypotheses. We showed encouraging preliminary results in terms of correlation with human judgement.

## 8. FUTURE WORK

We plan to explore other options for obtaining monolingual parallel data. One possible approach is to retrieve headlines of news articles from different agencies covering the same event. We expect headlines describing the same event to have some degree of semantic similarity yet different surface realizations.

The sequence-to-sequence model required a relatively long time to run which limited the testing of other architectures and model parameters. We plan to conduct more experiments on different architectures and compare results. More specifically, whether we get better results with a unidirectional model since sequences from the same language will tend to have similar word order. We also intend to incorporate the concept of “coverage” [16] to address issues with fluency of the neural model output.

We obtained encouraging results from the evaluation metric experiments but we need to verify its usefulness in settings with subtle semantic phenomena such as polysemy and synonymy. We also plan to use it at word subunits such as morphemes and even character level especially for surface distance comparison in morphology rich languages like Arabic.

## REFERENCES

- [1] Shady Abdel Ghaffar and Mohamed Fakhr. English to arabic statistical machine translation system improvements using preprocessing and arabic morphology analysis. Recent Researches in Mathematical Methods in Electrical Engineering and Computer Science, pages 50–54, 12 2011.
- [2] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, pages 19–27, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [3] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL'05, pages 597–604, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606, 2016.
- [5] Houda Bouamor, Aurélien Max, and Anne Vilnat. Comparison of paraphrase acquisition techniques on sentential paraphrases. In Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18, 2010, pages 67–78, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
- [6] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06, pages 17–24, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

- [7] Courtney Corley and Rada Mihalcea. Measuring the semantic similarity of texts. In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, EMSEE '05, pages 13–18, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [8] Michael Denkowski, Hassan Al-Haj, and Alon Lavie. Turker-assisted paraphrasing for english-arabic machine translation. <https://www.cs.cmu.edu/~mdenkows/pdf/paraphrasemturk-2010.pdf>, 2010.
- [9] Peter Wallis Dept and Peter Wallis. Information retrieval based on paraphrase. In In Proceedings of PACLING Conference, 1993.
- [10] Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada, and Masaaki Nagata. Learning to translate with multiple objectives. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 1–10, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [11] Rafael Ferreira, Rafael Dueire Lins, Fred Freitas, Steven J. Simske, and Marcelo Riss. A new sentence similarity assessment measure based on a three-layer sentence representation. In Proceedings of the 2014 ACM Symposium on Document Engineering, DocEng '14, pages 25–34, New York, NY, USA, 2014. ACM.
- [12] David Graff and Kevin Walker. Arabic newswire part 1, 2001.
- [13] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. <http://www.statmt.org/europarl/>, 2005.
- [14] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
- [15] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In Association for Computational Linguistics (ACL) System Demonstrations, pages 55–60, 2014.
- [16] Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. A coverage embedding model for neural machine translation. CoRR, abs/1605.03148, 2016.
- [17] Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Building a free, general-domain paraphrase database for japanese. In The 17th Oriental COCOSA Conference, Phuket, Thailand, September 2014.
- [18] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–51, 2003.
- [19] Ellie Pavlick and Chris Callison-Burch. Simple PPDB: A paraphrase database for simplification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers. The Association for Computer Linguistics, 2016.
- [20] Fabio Rinaldi, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. Exploiting paraphrases in a question answering system. In Proceedings of the Second International Workshop on Paraphrasing - Volume 16, PARAPHRASE '03, pages 25–32, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

- [21] Wael Salloum and Nizar Habash. Dialectal to standard arabic paraphrasing to improve arabic english statistical machine translation. In Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, DIALECTS '11, pages 10–21, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [22] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. CoRR, abs/1511.08198, 2015.
- [23] Ma Xiaoyi, Dalal Zakhary, and Moussa Bamba. Arabic news translation text part 1 ldc2004t17. <https://catalog.ldc.upenn.edu/LDC2004T17>, 2004.
- [24] Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. Paraeval: Using paraphrases to evaluate summaries automatically. In Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06, pages 447–454, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [25] Ingrid Zukerman, Bhavani Raskutti, and Yingying Wen. Experiments in query paraphrasing for information retrieval. In AI 2002: Advances in Artificial Intelligence: 15th Australian Joint Conference on Artificial Intelligence Canberra, Australia, December 2–6, 2002 Proceedings, pages 24–35, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg

*INTENTIONAL BLANK*



# APPLYING DISTRIBUTIONAL SEMANTICS TO ENHANCE CLASSIFYING EMOTIONS IN ARABIC TWEETS

Shahd Alharbi<sup>1</sup> and Dr. Matthew Purver

<sup>1</sup>Department of Software Engineering, King Saud University, Riyadh,  
Saudi Arabia

<sup>2</sup>School of Electronic Engineering and Computer Science, Queen Mary  
University of London, London, UK

## ABSTRACT

*Most of the recent researches have been carried out to analyse sentiment and emotions found in English texts, where few studies have been conducted on Arabic contents, which have been focused on analysing the sentiment as positive and negative, instead of the different emotions' classes. Therefore this paper has focused on analysing different six emotions' classes in Arabic contents, especially Arabic tweets which have unstructured nature that make it challenging task compared to the formal structured contents found in Arabic journals and books. On the other hand, the recent developments in the distributional semantic models, have encouraged testing the effect of the distributional measures on the classification process, which was not investigated by any other classification-related studies for analysing Arabic texts. As a result, the model has successfully improved the average accuracy to more than 86% using Support Vector Machine (SVM) compared to the different sentiments and emotions studies for classifying Arabic texts through the developed semi-supervised approach which has employed the contextual and the co-occurrence information from a large amount of unlabelled dataset. In addition to the different remarkable achieved results, the model has recorded a high average accuracy, 85.30%, after removing the labels from the unlabelled contextual information which was used in the labelled dataset during the classification process. Moreover, due to the unstructured nature of Twitter contents, a general set of pre-processing techniques for Arabic texts was found which has resulted in increasing the accuracy of the six emotions' classes to 85.95% while employing the contextual information from the unlabelled dataset.*

## KEYWORDS

*SVM, DSM, classifying, Arabic tweets, hashtags, emoticons, NLP & co-occurrence matrix.*

## 1. INTRODUCTION

The recent dominance of the electronic social media websites and especially Twitter which has become the major form of communication and expressing peoples' emotions and attitudes in the world generally and Arab world especially, has considered the significant encouraging factor for carrying this study. According to [1] which has published 2014 statistics for Twitter usage in Arab

world which showed increasing the number of active Twitter users in Arab world to more than 5 million. Additional remarkable statistic shows that the average number of tweets in 2014 written using Arabic language account over than 75% among other languages, which considered as a strong proof for the Arabic language as being one of the fastest growing languages on the web.

An additional encouraging factor, was based on the recent features and services provided by Twitter to its public users, such as streaming APIs supported by language mode which facilitate the automatic collection of a huge amount of its data without manual interventions, which has raised Twitter value as being a wealth source, gold mine and major area of interest to different researchers in the subjectivity analysis field to mainly analyse texts based on being subjective or objective.

Special area has been explored by researchers to analyse users' positive and negative sentiments and mine their opinions and reviews from unusual, and informal short texts which differ from the formal texts found in newspapers and documents.

Many of these researches have been applied to business tasks for companies to improve their services and products based on analysing customers' feedback. Similarly, customers can use analysed reviews to reveal insights about services and products. In addition, many other applications have been seen in determining public's attitudes and views with respect to special topic or incident.

The remaining sections of this paper is organised into a number of different sections, section 2 highlights some of the previous studies with their developed models in the different areas related to the distributional semantics and to the sentiment and emotions analysis; section 3 represents the methodology followed in building the proposed model for enhancing the automatic classification of the different emotions found in Arabic tweets, which was developed by [2]; in addition, section 4 describes some of the different carried experiments with their analysed results, and the final sections represents the conclusions drawn from the previously performed experiments as well as some of the faced limitations and the expected future work to eliminate these limitations.

## **2. RELATED WORK**

### **2.1. Sentiment and Emotion Analysis in Arabic Texts**

Lately the growth of the number of Arab social network users, produce a massive amount of Arabic texts and reviews available through the social medium which highlighted the need for automatic sentiment and emotions identification from this large generated reviews and texts.

Although some emotions and sentiment analysis studies have been conducted to non-English languages, most of these studies have been focused on non-Arabic languages. However, a sentiment analysis study was done by [3] to classify opinions in a number of web review forums, which was tested on Arabic dataset, through developing entropy weighted genetic algorithm, EWGA, for feature selection in addition to the different features extraction components that were used to compute Arabic features' linguistic characteristics. Results have proved the effectiveness of the built methodology for analysing and classifying sentiment in different languages through the use of SVM.

A former study was done by [4] which developed a tool dedicated for analysing colloquial Arabic texts that appear in the web forums and social media websites. A limitation of the dependency on human interventions and judgments to overcome the problems generated from the non-standardised colloquial Arabic texts has been resolved by their proposed tool through a game-based lexicon which is based on human expertise to classify the sentiment of the phrases.

Additional methodology was used for classifying sentiment, based on the subjectivity of the different words in the constructed lexicon of the words resulted from phrases segmentation.

It has been observed that many sentiment analysis studies have been conducted on analysing large texts and documents, however, few years after carrying out [4] study, [5] have analysed sentiments based on the sentimental majority in Arabic sentences, through calculating the number of positive and negative phrases and classify the sentence according to the dominant sentiment which recorded an accuracy of 60.5%. Their approach was based on their previously followed game-based approach to annotate large corpus manually.

Moreover, a model has been introduced by [6], known as SAMAR, which was designed to recognize the subjectivity and to identify sentiment for sentence-level Arabic texts in the social media websites' content such as tweets. Compared to [4], SAMAR has the ability to analyse sentiment found in MSA and colloquial Arabic phrases. To overcome the difficulty and the inaccuracy resulted from classifying sentiment in colloquial Arabic phrases that vary in the rules and vocabularies compared to MSA, analysed Twitter messages were annotated manually according to both, their subjectivity, as well as the different Arabic texts' classes as MSA and colloquial.

Despite the long success journey of analysing the sentiment and the subjectivity in Arabic texts, a lack of identifying emotions in Arabic contents of the social media was found which can lead to an inaccurate classification results for analysing emotion-based texts.

There was a remarkable solution developed by [2] which overcome this analysis limitation through classifying emotions found in Arabic tweets. This recent developed tool has the ability to detect the dialects found in Arabic tweets as well as to identify the different emotions used by [7] and others, through using SVM classification algorithm with n-gram order features and distant supervision approach similar to [7], [8] and [9] approaches. Results have recorded a good performance in predicting emotions for the different conventional markers in Arabic tweets according to the different emotions based on the human judgments, these results were between 93% and 74% and between 81% and 72% for emoticons and hashtags respectively.

## **2.2 Distributional Semantics**

It has been observed from the different studies in developing Distributional Semantic Model (DSM) approaches, the strong dependency on the assumption that the similarity of meanings between linguistic entities in any textual data can be defined in terms of the distributional properties of these linguistic entities. Therefore, the notion of utilizing these distributional properties for inferring meanings considered the hallmark of any developed DSM, which has come to be known as distributional hypothesis that considered the main idea behind the distributional semantics, therefore DSM need to be implemented on its basis. Rubenstein and Goodenough have reported on 1965 the early theory of distributional hypothesis in which words that have similar meanings occur in similar contexts. Based on their hypothesis, the predicted

similarities of the different words were evaluated based on their correlation with the predefined 65 noun pairs synonyms' similarities supported through the human judgments.

In the past four decades, a distributional model was developed by Harris who considered as the basic motivation for the distributional hypothesis where the differences in semantics between words can be correlated with the differences in distributions between these words.

Similar distributional hypothesis was defined by Schutze and Pedersen in 1995, such that different words will occur with similar neighbours if they have similar meanings based on the existence of sufficient text materials. Where the idea of the distributional semantics hypothesis defined by [10] is based on locating words with similar distributional properties in similar regions of the word-space.

Another hypothesis was defined by [11] model for testing the efficiency of the high order co-occurrence for detecting similarities. [12] have developed their approach based on their assumption that similarities in syntactic structure can result in semantic similarities.

Moreover, [13] model have depended on analysing Arabic texts similarities to identify the different participants groups presented in the online discussion based on the hypothesis that participants can share the same opinion in a discussion if they focus in the similar aspects of the discussion topic.

Despite the different distributional hypotheses which have been defined in different studies, they all fall under the idea that two words are expected to be semantically similar if they have similar co-occurrences in the observed context.

In addition to the distributional hypothesis, word-space vectors represents another fundamental basis for developing DSM approaches. In the past two decades, [14] found that DSM is based on the similarities and the distances between words in the vector space, in which semantically similar words are close, since semantics are expressed by the same set of words.

Moreover, [15] have found that word-space models utilises distributional patterns of phrases and words collected from large textual data to represent semantic similarities through the proximity between different phrases or words in an n-dimensional word-space.

Basically, two different approaches were recognized for the different developed models of DSM. The first methodology followed by researchers is based on the word-space vectors for building distributional profiles based on the surrounding words of the different terms as demonstrated by [14]. The second approach, known as latent semantic analysis model, LSA, which is one of the first applications of DSM and word-space vectors which are used in information and documents retrieval. This model is based on building distributional profiles of the words' occurrence based on the word-by-document matrix. [16] have introduced LSA approach using a word-by-document co-occurrence matrix in which contexts represented are based on the different documents in which words appear.

Despite the different types of distributional resources used for inferring similarities in these approaches, they share their main objective of inducing knowledge about meanings and similarities indirectly from co-occurrence of large textual data [10].

### 2.3. Semantic Similarity Detection in Arabic

It has been emerged a considerable number of studies for classifying Arabic texts and documents based on their similarities using distributional semantics and statistical properties with different similarity measures. This area has been seen as a successful application in the IR, text summarizing, document clustering, questions answering and other domains.

Based on [13] and their previously identified hypothesis, an approach has been developed to detect subjectivity of the Arabic discussions and to identify discussions' topics based on similarities between participants' opinions in online Arabic discussion groups. Similarities were detected using word-vector spaces of 100 dimension for representing different participants' discussions which contain distributional measures for the 100 different participated topics. Results have showed that the word-vector representations considered a rich representation for explicitly illustrating different discussions' topics. It has been observed the contribution of the topic representations and the distributional statistics on enhancing the accuracy of identifying subjectivity of the different Arabic discussions along with their topics.

According to [17], cosine-based similarity has considered one of the well-known similarity measures applied to the documents in different applications as IR, NLP, machine translation and text mining for Arabic documents classifications. [17] have found the effectiveness of using cosine similarity measures in Arabic document classification tasks compared to other distance-based and similarity measures.

Prior to [17], [18] have conducted a study which aims to find the optimal classification model for classifying Arabic texts. To find the best coefficient for the vector-space model to classify Arabic documents, these documents were represented as vectors to compare between the different coefficients, as cosine. Results have found the superiority of the cosine measure compared to other coefficients. Moreover, the effectiveness of applying cosine measure in classifying Arabic texts was compared to other similarity measures as Naïve Bayes which outperforms the used cosine-based similarity measure. [18] have suggested combining different classification models which can increase the classification accuracy for Arabic text, as Naïve Bayes with cosine measure to determine similarity in the classified text. Following [18], [19] have investigated determining the similarity between Arabic texts using different bigrams techniques, word-based model, document-level model and vector-based model which can reflect the importance of the bigrams in the document, based on the weighted vectors of the represented document. Cosine similarity measure has been applied to the vector-based model for measuring similarities between two vectors. On the other hand, word-based and document-level models have been used with different similarity measures as investigated by [20] for measuring similarities. Compared to the human similarity judgments, results have demonstrated the efficiency of the cosine similarity measures to determine similarities between Arabic texts and documents.

Although it can be observed from this set of highlighted studies the lack of classifying Arabic social media's texts based on applying different similarity measures, one of the most recent similarity and categorisation related studies was based on Arabic social media's texts especially Twitter posts, that was carried out by [21]. This study has delivered the problem of non-related results retrieved from searching Arabic tweets through developing a machine-learning model for summarising Arabic Twitter posts and especially Egyptian dialect posts which used cosine-similarity model as one of the developed approaches. A high performance was achieved by their proposed system compared to other summarisation algorithms.

## **2.4 Classification and Annotation**

### **2.4.1 Arabic Text Classification**

A considerable number of Arabic texts classification studies have used a similar classification methodology which was used in [7], [8] and [22] tools for classifying Arabic sentiments and emotions found in the different tweets.

Following [7], [2] have developed a tool which was built using SVM classifier for detecting and classifying emotions and dialects in Arabic tweets based on the same conventional markers and emotions' classes used by [7]. It has been proven that following [7] and [22] approaches of using distant supervision classification approach to automatically label emotions and dialects in Arabic tweets, can achieve more reliable results for emotions and dialects classification accuracies. Results have showed that changes in the classifier accuracies achieved, were based on the application domain (emotion or dialect detection). This has supported [23] findings regarding the reliance of the machine-learning classifiers' performance on the domain applications used during the training process.

A study was carried out by [24] which encouraged the use of SVM in the area of classifying and categorizing Arabic texts, consequently, a comparative study was conducted by [25] for comparing between two machine-learning classification approaches including SVM for categorising Arabic texts using a large number of training and testing articles. Results have showed the role of features set size on the SVM performance, as it has been observed that larger set of features has increased SVM classification outcomes. A similar classifiers' comparison based study was done by [26] based on applying SVM and other traditional machine-learning classification techniques as Naïve Bayes classifier in classifying Arabic texts. Results have supported [25] findings, in which a better performance from SVM classifiers was generated when increasing the features set size, due to its ability to deal with sparse vectors of the classified documents. Moreover, the high dimensions of the vector-space represented from classified dataset enable SVM in handling large number of features.

Alternatively, many remarkable Arabic text classification studies have used different classification techniques. [27] has compared the efficiency of using distance-based and different machine-learning techniques as Naïve Bayes to classify large Arabic documents represented using word-vector spaces with their frequencies. Results have showed that Naïve Bayes machine-learning model has outperforms the distance-based model for classifying Arabic texts.

## **2.5 Arabic Text Pre-processing Approaches**

Based on the different Arabic language specifications, many Arabic text classification studies have investigated the effect of the language properties on the trained and the classified texts. A study was carried by [28] which explored the effect of different Arabic pre-processing techniques on classifying texts by applying different term weighting and stemming approaches. According to [28] experiments, result have showed the superiority of SVM to classify Arabic processed texts compared to other text classifiers, moreover it has been found that the light stemming, considered the best feature reduction technique. Similarly, [29] have investigated the impact of using stemming in the Arabic text pre-processing. Results have showed that using stemming, SVM accuracy was increased compared to Naïve Bayes classifier. Moreover, [30] model has pre-

processed the review dataset using different approaches to remove spammed, noisy and duplicated reviews to avoid inconsistency during processing reviews and to guarantee a unique dataset contents which was used in the three different dictionaries, Arabic, English and emotions. Arabic opinion analysis model proposed by [31], has performed different pre-processing schemas as removing punctuations, symbols, digits from dataset, applying tokenisation as well as filtering non-Arabic texts to normalize analysed opinions. Recent development of Arabic text classification that was built by [2] for classifying Arabic emotions and dialects in Twitter messages has investigated the impact of applying similar pre-processing techniques that was followed by [28]. In addition to the different pre-processing schemas [2] and [28] have investigated the impact of normalising Arabic text as well as removing stop words as a feature reduction. Results examined by [2] have showed the correlation between SVM classification accuracy and different pre-processing techniques applied to classify emotions in Arabic tweets.

On the other hand, [32] have pre-processed Arabic documents that were used in the keyword extraction system to reduce features by portioning documents into sentences and removing the non-candidate keywords. [26] model has used SVM to classify Arabic texts with features' selection schemas during pre-processing steps. Stemming and eliminating stop words were used to reduce dimensionality. Results showed the positive effect of the light stemming on the Arabic text classification.

### **3. METHODOLOGY**

The methodological approach followed in this study was based on applying the DSM measures from the unlabelled collected random dataset to build the co-occurrence matrix which was used to derive features that encode meaning and similarity. These features were then added into the standard n-gram features' order representation used in [2] classification approach in order to prove our early stated hypotheses for their ability in enhancing [2] model for classifying emotions found in Arabic tweets. The process followed was mainly divided into four stages, preparing labelled (keyword) and unlabelled (random) datasets, measuring co-occurrences for unlabelled dataset terms, preparing the best pre-processing techniques which was applied in [2] model and building the feature vectors for our classified datasets. For classification improvement purposes, the outcome of the latter stage was then used with the SVM classifier model developed and used by [2].

#### **3.1 Preparing the labelled and unlabeled datasets**

Two different Arabic Twitter datasets, labelled and unlabelled datasets, were collected automatically using some available Java libraries, Twitter4j. Streaming API was the main Twitter API used to access a global stream of random Arabic tweets.

The collected labelled dataset was used during the training and the testing stages of the classification process, with the different conventional markers, hashtags and emoticons, in the collected posts for the different six emotions' classes, happy, sad, anger, fear surprise and disgust. Where the collected unlabelled dataset was used for building the co-occurrence matrix which represented the contextual information. In both datasets a language request parameter was set to Arabic for restricting the collected data. Both datasets were further processed to eliminate duplicate and retweet data used during the classification process to avoid bias results in both datasets. Additional process included for the labelled dataset to eliminate tweets which include

mixed labels from different emotions' classes, to avoid confusing the classifier during the classification tasks with mixed labels in the training and testing sets. This features' reduction process applied at early stages considered a required step which was performed before classifying the different emotions. On the other hand, pre-processing techniques considered an optional features' reduction techniques which were applied when required.

### **3.2 Measuring co-occurrences for the unlabelled dataset**

As DSM is based on obtaining distributional information in a high dimensional vectors, the proposed model represents these distributional information in  $R \times C$  matrix to capture the co-occurrence frequencies measures. Although, it has been proven from previous studies the ability of the co-occurrence statistics to provide semantic information, the proposed approach aims to reveal an extension for its ability to strengthen the classification job by exploiting these measures.

### **3.3 Pre-processing Arabic Text**

According to the Arabic text classification studies which were highlighted earlier, Arabic text-pre-processing techniques have proved their effective impact on the classification task. In our model this effect have a special importance during the feature reduction process due to the noisy and unstructured nature of the Arabic texts produced in Twitter by different users which makes similar messages undistinguishable by the classifier due to the differences in their presentation. Consequently, the proposed model has applied some optional techniques to normalise Arabic tweets for the classifier to simplify its job and therefor better results can be achieved. Although there was an absence for a standard defined set of techniques for pre- processing Arabic texts, our model has applied [2] approach using six different techniques during this optional feature reduction step such as removing the Arabic stop words, stemming, normalisation, removing diacritics as well as lengthening characters and reducing repeated characters. Unlike [2] during this processing stage our model has applied their techniques on the vocabularies-level for co-occurrence features as well as the tweets-level for the labelled dataset.

Choosing the best number of co-occurrence dimensions was based on the best average classification accuracy for the different emotions resulted from the different tested dimensions. Similarly, the best set of pre-processing techniques was selected based on the highest accuracy average resulted from classifying the six emotions labelled using hashtags and emoticons.

### **3.4 Building Features' Vectors**

Apparently, co-occurrence statistics and distributional measures obtained from the co-occurrence matrix solely cannot considered as sufficient representations to classify different emotions occurred in the Arabic tweets, according to [18] finding. [33] have suggested adding semantic features to improve the identification of Twitter sentiments, on the other hand, theproposed model was based on adding the co-occurrence measures, correspond to the contextual information, as additional vectors to support the features' vectors with similar measures for emotions which were expressed using similar features in the classified tweets, which can revel similarity between emotions and therefore help the classifier to distinguish easily between emotions.



The result from this process for both datasets was a collection of features' vectors with their term frequencies. In addition to the similarities between featurising both datasets, few differences in constructing these vectors were encountered.

Following [2] featurisation methodology, labelled features' vectors were generated based on the different terms that were separated by whitespaces in the labelled dataset tweets which were used in the training and the testing sets. On the other hand, features' vectors from the unlabelled set were generated from R normalised rows of C dimensions in the RxC co-occurrence matrix that was built from the unlabelled dataset tweets regardless of the labelled tweets used in the training and learning dataset. These features' vectors consists of frequencies' probabilities between different vocabularies occurred in the unlabelled tweets, where features' vectors from labelled dataset were based on the terms' frequencies of the tweets' terms used during the training and testing processes.

After generating the labelled features' vectors set  $\{kV1, kV2, kV3, \dots, kVz\}$  from the labelled tweets, an equal number of co-occurrence features' vectors set,  $\{rV1, rV2, rV3, \dots, rVz\}$ , were generated. For each labelled feature vector  $kVi$  a new feature vector,  $kVi'$ , was generated to be used during the classification process which substituted vector  $kVi$ . This new generated vector  $kVi'$  was resulted from appending a labelled feature vector for each tweet  $kVi$  with the unlabelled feature vector  $rVi$  that was obtained from combining different rows' co-occurrence vectors  $kfj$  of C dimensions of every feature which occurred in  $kVi$ , therefor C dimensions were consistent among all the calculated co-occurrence features' vectors,  $rVi$ , as well as the co-occurrence features' vectors,  $kfj$ , used in generating  $rVi$ , on the other hand the number of dimensions for each keyword features' vectors,  $kVi$  and therefore the final resulted  $kVi'$ , was various for each tweet used in the training and testing process, this variation was based on the number of features generated during the tokenisation process for the classified tweets in the keyword dataset.

Figure 3 1 simplifies the methodology used as well as the featurisation process followed in the proposed model, which is repeated for each tweet used in the training and testing set, in our case this process was repeated 500 times as we used  $N=500$  in each experiment.

Similar to [2] model, the labelled features were generated using n-grams features order, however, co-occurrence features' vectors  $rVi$  appended to the labelled features' vectors  $kVi$  were based on the occurrence vector for each unigrams features only,  $kfj$  in  $kVi$ . These features' vectors composed  $kVi'$  were used by the SVM classifier which is explained in the following section.

### 3.5 Classifying Arabic Tweets in the Labelled Dataset

Throughout all the following experiments which the model have performed, SVM classifier was used to classify emotions in different Arabic tweets. In this emotions' classification model, the main goal of the SVM is to discriminate the rule used during the learning process for separating the different emotions accurately into positive and negative set based on the target emotion class with an optimal separating hyper-plane to attain the minimum error rates over the target emotions' classes. Emotions in this model were classified using the distant supervision learning approach, through predicting a predefined emotions classes based on the labels, hashtags and emoticons, used in annotating the labelled dataset, which were then removed from the classified emotional tweets as well as using the co-occurrence information of the unlabelled data in order to help improve the classification performance.

The proposed model has followed [7] and specially [2] classification approach for classifying the six emotions used in Arabic Twitter messages which was based on SVM with the support of both LibSVM and LibLINEAR libraries. Based on [26] proof of the correlation between the features set and the SVM classification accuracy, this has showed the ability of the SVM in handling a large number of features' vectors and dimensional data. Similar to [7], [2], [29] and [34], classification accuracy was based on the K-fold-cross-validation technique, as [2] and [7] the model used 10-fold-cross-validation, in which the classified data set was divided into 10 equal parts, 9 parts out of the total parts were used for the training purposes while the last part was used for the testing purposes. To ensure accurate result all parts need to be tested, through repeating the process 10 different times where the final accuracy result was the average of the 10 repeated processes.

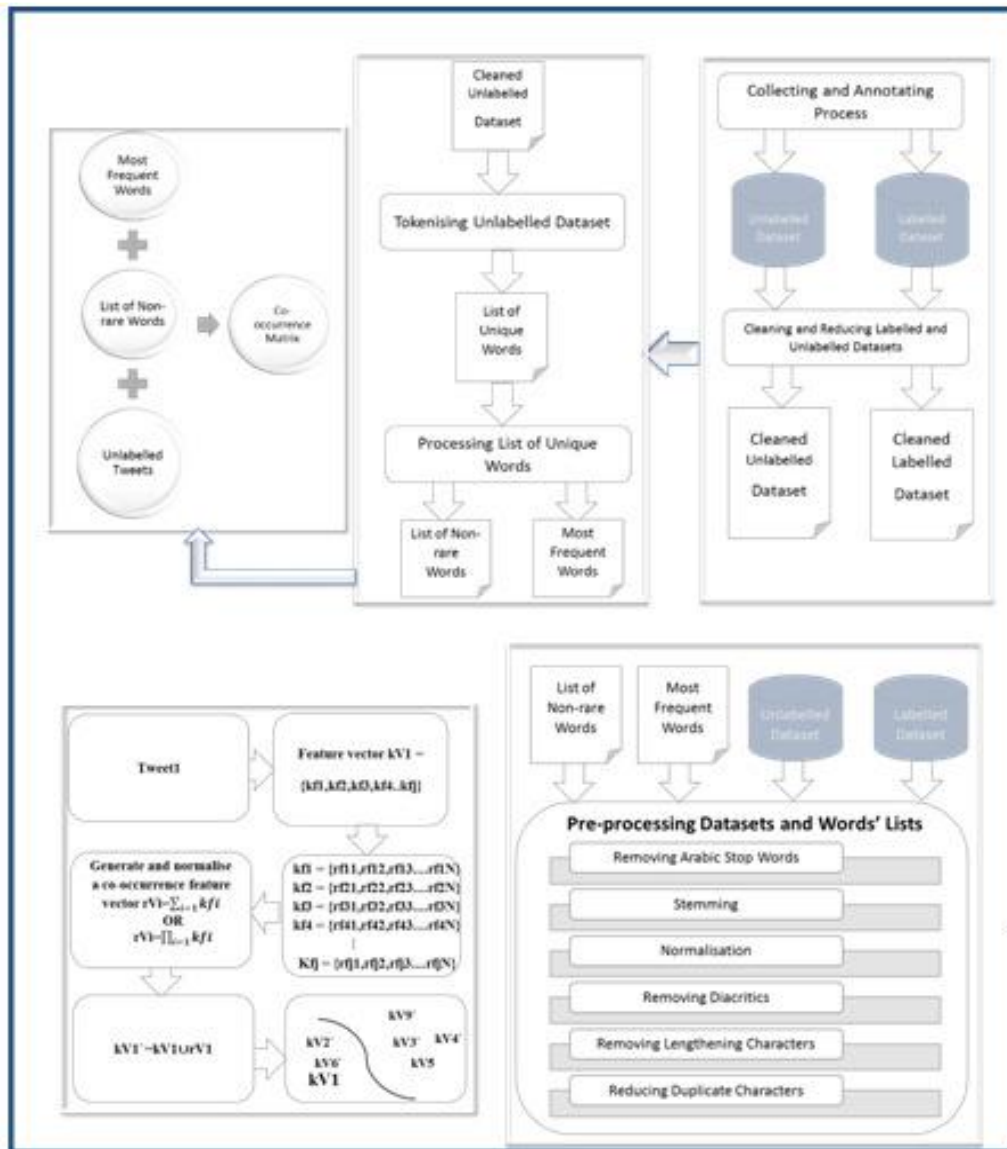


Figure 3 1 Model methodology

## 4. DATA

Table 4-1 Total unlabeled dataset

UNLABELLED DATASET	
Total	1,368,849 Tweets

Table 0-1 Total labelled dataset

### BASE LABELLED DATASET

Happy Hashtag	Sad Hashtag	Anger Hashtag	Fear Hashtag	Surprise Hashtag	Disgust Hashtag
2,758	1,697	489	946	578	628
Total			7,096 Tweets		
Happy Emoticon	Sad Emoticon	Anger Emoticon	Fear Emoticon	Surprise Emoticon	Disgust Emoticon
103,741	21,322	2,381	4,147	3,068	12,878
Total			147,537 Tweets		

## 5. EXPERIMENTS AND RESULTS

During this study a number of experiments was carried out to prove our hypothesis through investigating the effect of adding the contextual information of an unlabelled dataset to [2] model, using the proposed methodology from the previous section.

In general two different set of experiments were performed, preliminary experiments as well as a number of underlying experiments based on applying the best techniques investigated during the first set of experiments.

The introductory experiment within the basic set of experiments aimed to test the ability of the built co-occurrence matrix in detecting the similarities between different words and emoticons which exist in the unlabelled dataset. Accordingly, based on the achieved results for the ability of the developed co-occurrence matrix to induce similarities, our methodology was applied to carry the other set of experiments and investigate the effect of the unlabelled dataset on enhancing the classification task, therefore, the following two experiments have explored the effect of both the dimensionality as well as the number of the unlabelled co-occurrence vocabularies used in the co-occurrence matrix on the classifier performance. Following these two experiments, the classifier accuracy was tested with different co-occurrence matrix types binary and frequency co-occurrence matrix. Moreover, as illustrated in the featurisation process, different generated

vectors from the co-occurrence matrix were integrated using different approaches additive and multiplicative approaches, their effect on the classification task was investigated in the fifth experiment of the basic set of experiments. An additional experiment was based on the labelled dataset only without employing our model in including contextual information during the classification process, this experiment was carried out to unify the preprocessing techniques to be applied for the six different emotions with both labels, instead of using different pre-processing schema for each emotion. As a result, the outcomes from these six preliminary experiments were used as the basic features for our developed model to carry out the four additional underlying experiments. After determining the basic model specifications, experiments within the second set were executed to investigate the impact of the previously determined pre-processing techniques set on both datasets for classifying different emotions with the both hashtags and emoticons labels. Furthermore, as Arabic stop words considered the most common occurrence words which do not convey extra meanings, our model has examined their effect by testing the contribution of their contextual information in the classification task before and after removing these words. And finally, the effect of increasing both the labelled dataset as well as the unlabelled dataset was investigated, which was used in the co-occurrence matrix, on the classifier performance to detect different emotions. Some of these experiments are highlighted in this paper.

To guarantee an equal number of positive and negative tested and trained tweets throughout all the different carried experiments and therefore to avoid bias results for popular emotions, the model was tested on a dataset with 500 tweets,  $N=500$ , this considered a larger dataset compared to [2] model which used 300 tweets. The selected dataset was divided between positive and negative classes for each experiments equally,  $N/2$  tweets were assigned for each positive and negative class. For the negative class  $N/2$  was divided equally between the other emotions except the target emotion which already has been assigned to  $N/2$  tweets labelled with the target emotion and conventional marker.

### 5.1 Testing the Effect of the Co-occurrence Matrix Dimensions

The aim of this experiment was to determine the most optimum number of column's and row's dimensions used in the co-occurrence matrix. The decision was based on testing the impact of these dimensions on the accuracy of classifying the different emotions occurred in Arabic tweets of our labelled dataset. In order to confirm the optimum number of dimensions, we have gradually increased the dimensions to measure their effect accurately, where in [32] methodology the co-occurrence dimensions were set directly to the most frequent 10 terms.

chart 5-1 and chart 5-2 illustrate the correlation between the different number of dimensions and the average accuracy for classifying the six emotions with both conventional markers, hashtags and emoticons compared to the accuracy which was resulted from [2] model for classifying these six emotions without employing co-occurrence information from our unlabelled random dataset which was considered as the baseline accuracy throughout the different carried experiments. Regardless to the different number of dimensions tested in this experiment, it was clear from the early results the contribution of including contextual information with different classified features' vectors form the labelled dataset in increasing the accuracy of classifying the different emotions included in the Arabic tweets of our labelled dataset, as increasing these dimensions resulted in adding more contextual information for each feature which occurred in the features' vectors of the different positive and negative tested and trained tweets during the classification process. Although this increase has remained almost steady with the different tested dimensions, it has been proven that including the contextual information of our unlabelled dataset not only

helps in the classification task, but also makes a quite high differences in the classification accuracy, as the average increase was more than +8.5% and +9.53% using different number of the column's and row's dimensions respectively.

Although it is clear the considerable difference between the baselines averages, these differences have been reduced to +0.39% and +0.57% when using our proposed model for including contextual information from our random dataset using different row's and column's dimensions to the classified emotions included in different keyword dataset tweets.

According to the proposed results, 1,000 column dimensions have achieved the highest average accuracy, although 5,000 dimensions were chosen as the optimum column's dimensions from the set of the 100, 500, 1,000, 2,000 and 5,000 tested dimensions to build our co-occurrence matrix. This choice was supported by our belief that the more contextual information we include from the co-occurrence matrix as column's dimensions, the higher accuracy we can achieve. Moreover, as we have built the co-occurrence matrix using 10,000 random tweets only, generalising these 1,000 dimensions as the optimum number of dimensions can be considered a risky decision, as the first 10,000 random tweets may fail to represent their strong correlations with the different tested numbers of dimensions. In addition to these reasons, the minor difference, 0.08%, observed from the average accuracies for detecting the different emotions with 1,000 and 5,000 dimensions have provide a further justification to consider 5,000 as the base column's dimensions in our built co-occurrence matrix instead of 1,000 dimensions.

On the other hand, 10,000 vocabularies have been selected for building the constructed cooccurrence matrix. Although 1,000 and 5,000 vocabularies have recorded higher accuracies compared with the classification performance when using 10,000 vocabularies, this selection was based on our aim to maximise the number of co-occurrence feature vectors, denoted as  $kf_1, kf_2 \dots kf_j$  in Figure 3-1, for every feature occurred in the positive and negative tweets used during the classification process, hence, this can be considered as an inevitable trade-off between the small differences of the achieved accuracies and increasing the number of co-occurrence feature vectors used during the classification and featurisation processes.

On the other hand although 30,000 co-occurrence vocabularies include more vocabularies compared to 10,000 co-occurrence vocabularies, this choice was eliminated as it has been observed from the built co-occurrence matrix a six vocabularies, 0.02%, out of the selected 30,000 did not have any occurrence associations with the previously selected 5,000 dimensions where with the 10,000 vocabularies a 100% associations were found between the vocabularies and the 5,000 co-occurrence matrix dimensions.

As a concluding result, we have observed that by just including a contextual information from 10,000 unlabelled tweets, classifying different emotions has been improved, this has also proved that the information captured from e.g. 100 dimensions and 1,000 co-occurrence vocabularies can be quite enough to produce high performance improvements, which indicates that contextual information gained from adding more hundreds or even thousands dimensions and/or co-occurrence vocabularies does not generate huge differences. Although this cannot be generalised as our contextual information was captured from 10,000 unlabelled random tweets only.

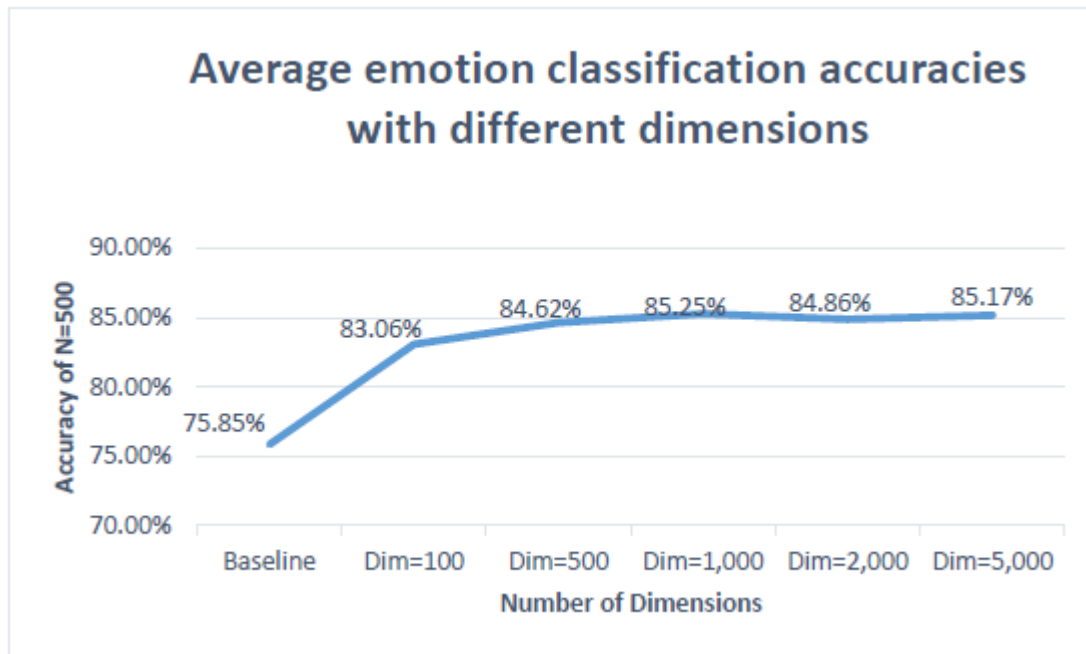


chart 5-1 Co-occurrence matrix effect on the classification accuracy using different column's dimensions

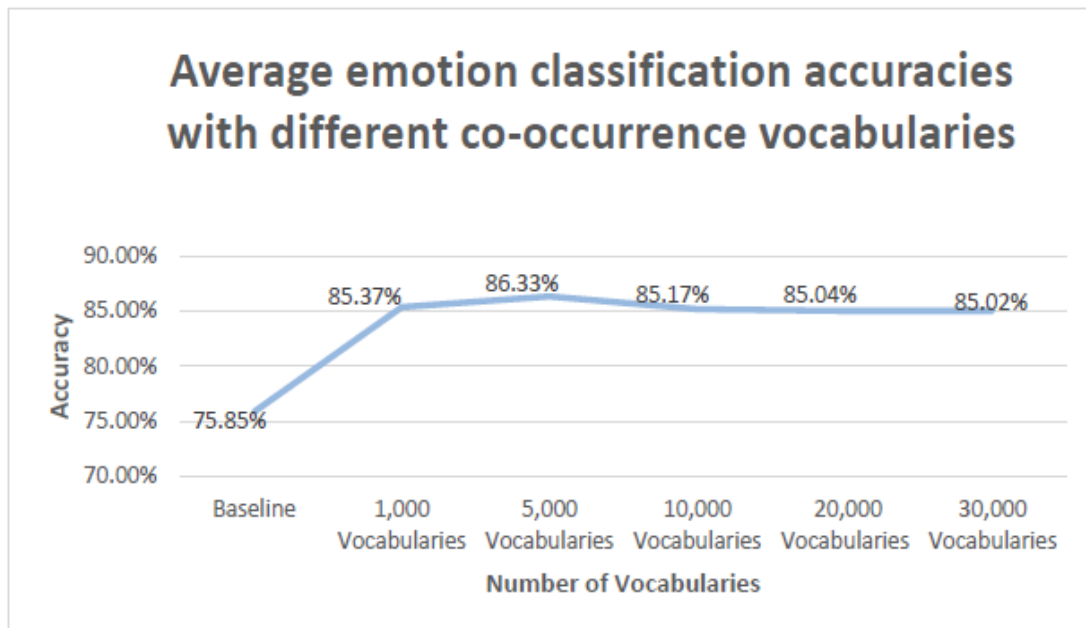


chart 5-2 Co-occurrence matrix effect on the classification accuracy using different row's dimensions

## 5.2 Removing emotions' labels from the co-occurrence column's dimensions in the co-occurrence matrix

This experiment has answered two questions which were concerned about measuring the popularity of the different labels used in annotating tweets from the labelled dataset as well as the amount of the contextual information these labels can provide.

It has been observed during this experiment that some emoticons have been presented as dimensions in the previously chosen 5,000 dimensions, on the other hand, these dimensions have showed the absence of the different used hashtags for the six target emotions. This can indicate the dominant of the emoticons in Arabic tweets compared to the emotional hashtags which was already proved from the total number of collected tweets using hashtags and emoticons for the six different emotions as shown in Table 4 2. Moreover, as emotional hashtags are typed expressions, different users can express the same emotion using different hashtags e.g. (#فرح) and (#فرررح), (#happy) and (#haaaappy) are different hashtags for the same emotion generated from the same adjective (فرح), (happy). Consequently, this has led hashtags to be less probable to appear in our most frequent 5,000 column's dimensions. Therefore this experiment has been performed to the six different emotions labelled with emoticons only.

From the 5,000 column's dimensions only 8 dimensions have been corresponded to different emoticons with a probability of 0.0016% , where 7 of these 8 emoticons [:(, :D, (:, :\$, :-D, ;), =D, :p], belong to happy class.

Since emoticons are being used essentially as labels in the labelled dataset, using these labels' contextual information during the classification process which explicitly encode the co-occurrence with those same emoticons can generate a bias classification results. Therefore checking that removing the emoticons co-occurrence information doesn't damage the performance is considered an important task which was checked in this experiment through removing the 8 emoticons from the 5,000 dimensions. Consequently, a slight classification decrease has been observed, for classifying emotions with both conventional markers, hashtags and emoticons, from 85.60% in the presence of these 8 emoticons as column's dimensions in the co-occurrence matrix to 85.30% after removing these emoticons. chart 5 3 illustrates the changes between the two approaches compared to the baseline. Although this performance has been dropped compared to the performance when labels information was included in the co-occurrence matrix, the model has continued to improve the classifier performance for all the tested emotions, except for sad and anger, even with the absence of the labels' contextual information from the unlabelled dataset presented as column's dimensions in the co-occurrence matrix.

Therefore, the presented results can prove the strong contribution of the co-occurrence information included during the classification task, in detecting emotions even in the absence of the information associated with the emotions' labels, e.g. emoticons. This strong contribution found is due to the ability for the co-occurrence information to capture related contextual data from the unlabelled tweets which is associated with the different features included in the labelled dataset features' vectors of the classified positive and negative tweets.

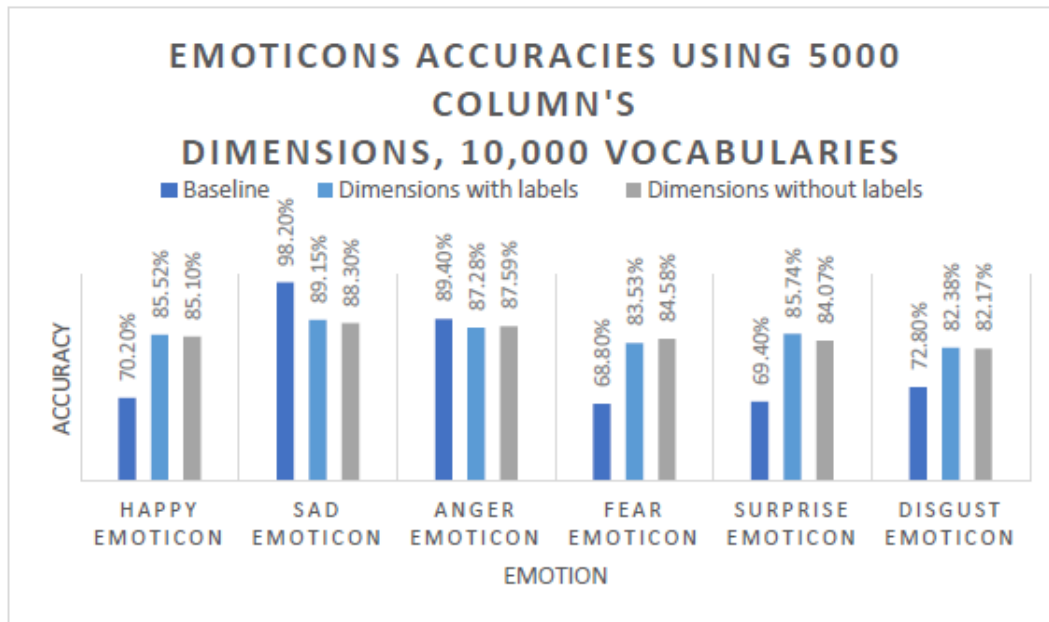


chart 5-3 Emotion classification accuracies using different approaches

### 5.3 Pre-processing Labelled (Keyword) and Unlabelled (Random) Datasets

This experiment was carried out to investigate the impact of applying the best pre-processing techniques on both datasets, which were inferred from one of the experiments on the labelled dataset. This impact was investigated based on the classification accuracies' changes when co-occurrence information was included during the classification process. The selected set of pre-processing techniques has increased the average performance for classifying hashtags and emoticons from 75.85%, as a baseline average accuracy, to 81.17%. Moreover, this average performance was improved to 85.95% when including pre-processed contextual information of the unlabelled dataset to classify pre-processed tweets from our labelled dataset through applying the best induced techniques.

As illustrated in chart 5 4, applying pre-processing techniques to the different positive and negative labelled dataset and to the co-occurrence information of the unlabelled dataset used in the classification process, have increased the accuracy for classifying hashtags and emoticons with an average increase of +5.02% and + 4.55%. This increase has highlighted the positive impact of compositing the co-occurrence and contextual information of the features with similar contexts under one common feature, on the classification performance, since applying the chosen set of pre-processing techniques have produced common words from different set of words, e.g. in case of applying stemming to the unlabelled dataset, the contextual information of (حزني), (my sadness) and (الحنن) (sadness) combined under (حزن), (sad), similarly the same is applied to the reduction of the repeated characters technique as well as using normalisation technique. Therefore, in the case of pre-processing both datasets, each feature's vector generated from the co-occurrence matrix is a result of combining features' vectors of similar words, which therefore resulting in combining contextual information of similar contexts, referring to the previous example, in case of having (حزني), (my sadness) or (الحنن) (sadness), or (حزن), (sad) as features in



the labelled features' vectors used during the classification, the unlabelled feature vector of the stemmed feature (حزن), (sad) will be used in all the different cases where (حزني), (my sadness) or (الْحزن) (sadness), or (حزن), (sad) occurred in the classified tweets. Consequently, this feature vector of the stemmed feature (حزن), (sad) captured the contextual information of all the features originated from this stemmed feature, instead of treating them independently. Although these different pre-processing techniques, except for removing stop words technique, have reduced both the number of row's and column's dimensions of the co-occurrence matrix as they generate number of similar words which were illuminated, their positive impact described previously have prevailed the impact of losing number of repeated dimensions.

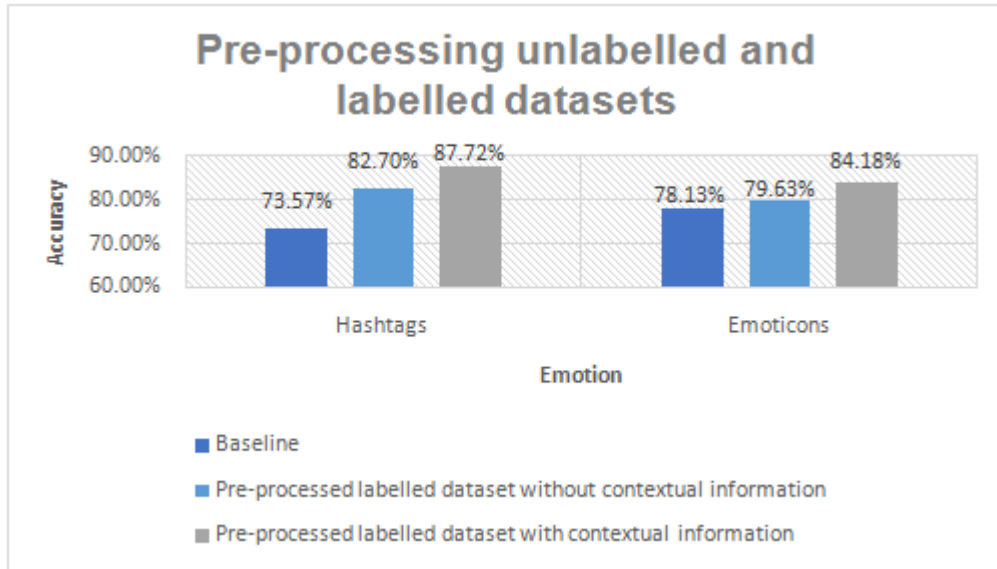


chart 5-4 Classification accuracies before and after pre-process tweets

## 6. CONCLUSION

During this study a semi-supervised learning approach has been developed to test the effect of the DSM, co-occurrence and contextual information statistics on enhancing the performance of [2] model developed using SVM classifier for the automatic detection of the different emotions' classes found in Arabic tweets collected using hashtags and emoticons for the six target emotions. This has resulted in higher accuracies compared to the standard supervised learning tasks and [2] model, through increasing the amount of the classified texts to include contextual information collected from unlabelled dataset instead of including a large amount of labelled sources, which can be considered inefficient way of increasing the accuracy, due to the time and cost limitations that can be faced in some situations.

Moreover, the model has tested the similarity between different emoticons using co-occurrence information, which proved the sensibility of the captured contextual information. This similarity test has also revealed the sensibility use of the emoticons found in Arabic tweets in relation to the descriptive texts surrounding the different emoticons.

During all the different investigated cases, the model has successfully achieved higher accuracy percentages compared to [2] model when classifying the different hashtags and emoticons for the six emotions, these increases varies based on the different tested factors. Consequently, our experimental results have showed a significant increases in the classification averages for both hashtags and emoticons, which indicate the positive impact of including contextual information during the classification process.

Therefore, based on these achieved results, our model has recorded a higher average accuracy for the six emotions' classes which was more than 86%, compared to most of the highlighted studies which analysed sentiment and emotions in Arabic texts, as illustrated in Table 6 1. On the other hand, a higher accuracy was achieved by [35] which was based on a constructed sentiment dictionary as highlighted previously. Moreover, [3] model have recorded a 95% accuracy which was based on a using a feature selection and extraction algorithm to compute Arabic features' linguistic characteristics, while the developed model did not depend on knowledge sources other than the contextual information of the unlabelled collected tweets.

Table 6-1 The achieved accuracy for some Arabic sentiment and emotion analysis studies

Arabic Sentiment/Emotion Analysis Model	Base/Average Accuracy Achieved
[30]	54% using SVM
[5]	60.5%
[6]	65.87%
[2]	72%

## 7. FUTURE WORK

As the constructed co-occurrence matrix was built to capture contextual similarities between two different words, contextual information from the unlabelled dataset of bigrams and n-gram features' order included in the labelled set was not included during the classification process. Therefore as a future work, we are planning to extend the co-occurrence matrix to capture their contextual information, as well as to capture the contextual for a larger number of unlabelled tweets to build the co-occurrence matrix. Moreover, we are planning to reflect the effect of different Arabic negation terms in the built co-occurrence matrix. Succeeding [2] model for automating the detection of the different Arabic dialect, we are also planning to extend enhancing the detection of the different emotions on a dialect level using DSM features.

## REFERENCES

- [1] ASMR, 2014 . Citizen Engagement and Public Services in the Arab World: The Potential of Social Media, Dubai: Mohammed Bin Rashid School of Government.
- [2] AlMutawa, B. & Purver, M., 2013. Automatic emotion and dialect detection tool for Arabic language, London : Queen Mary University of London.
- [3] Abbasi, A. et al., 2008 . Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Transactions on Information Systems (TOIS), 26(3), pp. 12-46.
- [4] AlSubaihin, A. et al., 2011. A proposed sentiment analysis tool for modern Arabic using human-based computing. New York, ACM.

- [5] AlSubaihin, A. & AlKhalifa, H., 2014. A System for Sentiment Analysis of Colloquial Arabic Using Human Computation. *The Scientific World Journal*, 2014(2014).
- [6] Abdul-Mageed, M. et al., 2012. SAMAR: a system for subjectivity and sentiment analysis of Arabic social media. PA, USA, Association for Computational Linguistics.
- [7] Purver, M. & Battersby, S., 2012. Experimenting with distant supervision for emotion classification. Stroudsburg, Association for Computational Linguistics, pp. 482-491 .
- [8] Yuan, Z. & Purver, M., 2012. Predicting Emotion Labels for Chinese Microblog Texts. London, Proceedings of the ECML-PKDD 2012 Workshop on Sentiment Discovery from Affective Data (SDAD).
- [9] Plutchik, R., 1980. A general psychoevolutionary theory of emotion. In: *Emotion Theory, Research, and Experience*. New York: Academic Press.
- [10] Sahlgren, M., 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1), pp. 33-54.
- [11] Wiratunga, N. e. a., 2007. *Acquiring Word Similarities with Higher Order Association Mining*. Berlin, Springer.
- [12] Fürstenau, H. & Lapata, M., 2012. Semi-supervised semantic role labeling via structural alignment. *Computational Linguistics*, 38(1), pp. 135-171 .
- [13] Abu-Jbara, A. et al., 2013. Identifying Opinion Subgroups in Arabic Online Discussions. s.l., Proceedings of ACL.
- [14] Schutze, H. , 1992. *Dimensions of Meaning*. Minneapolis, Proceeding of Supercomputing.
- [15] Sahlgren, M., 2006. *The Word-Space Model* , Sweden: University of Stockholm.
- [16] Deerwester, S. et al., 1990. Indexing by Latent Semantic Analysis. *THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6), pp. 91-407.
- [17] Froud, H. et al., 2013. Arabic Text Summarization Based on Latent Semantic Analysis to Enhance Arabic Documents Clustering. *International Journal of Data Mining & Knowledge Management Process (IJDMP)*, 3(1), pp. 79-84.
- [18] AlKabi, M. & AlSinjilawi, S., 2007. A Comparative Study of the Efficiency of Different Measures to Classify Arabic Text. *University of Sharjah Journal of Pure and Applied Sciences*, 4(2), p. 13 – 24.
- [19] AlRamahi, M. & Mustafa, S., 2011. N-Gram-Based Techniques for Arabic Text Document Matching; Case Study: Courses Accreditation. *ABHATH AL-YARMOUK*, 21(1.), pp. 85-105.
- [20] Khreisat, L. , 2006. Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. Las Vegas, Proceedings of the 2006 International Conference on Data Mining, DMIN .
- [21] ElFishawy, N. et al., 2014. Arabic summarization in Twitter social network. *Ain Shams Engineering Journal*, 5(2), p. 411–420.
- [22] Go, A. et al., 2009. *Twitter Sentiment Classification using Distant Supervision*. Project Report, Stanford, p. 1–12.

- [23] Taboada, M. et al., 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37 (2), pp. 267-307.
- [24] Mesleh, A., 2007. Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. *Journal of Computer Science*, 3(6), pp. 430-435.
- [25] Hmeidi, I et al., 2008. Performance of KNN and SVM classifiers on full word Arabic articles. *Advanced Engineering Informatics*, 22(1), pp. 106-111 .
- [26] Gharib, T. et al., 2009. Arabic Text Classification Using Support Vector Machines. *International Journal of Computers and Their Applications*, 16(4), pp. 192-199.
- [27] Duwairi, R., 2007. Arabic Text Categorization. *The International Arab Journal of Information Technology*, 4(2), pp. 125-131.
- [28] Saad, M. , 2010. *The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification*, Gaza: The Islamic University.
- [29] Rushdi-Saleh, M. et al., 2011. OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, 62 (10), pp. 2045-2054.
- [30] AlKabi, M., AlQudah N. et al., 2013. Arabic / English Sentiment Analysis: An Empirical Study. Irbid, The 4th International Conference on Information and Communication Systems (ICICS).
- [31] AlKabi, M. et al., 2014. Opinion Mining and Analysis for Arabic Language. (IJACSA) *International Journal of Advanced Computer Science and Applications*, 5(5), pp. 181-195.
- [32] AlKabi, M., AlBelaili, H. et al., 2013. Keyword Extraction Based on Word Co-Occurrence Statistical Information for Arabic Text. *ABHATH AL-YARMOUK*, 22(1), pp. 75- 95.
- [33] Saif, H. et al., 2012. *Semantic sentiment analysis of twitter*. Heidelberg , Springer-Verlag, pp. 508-524 .
- [34] Chaovalit, P. & Zhou, L., 2005. *Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches*. Washington, IEEE Computer Society.
- [35] AlKabi, M. et al., 2013. *An Opinion Analysis Tool for Colloquial and Standard Arabic*. New York, ACM.

# GENERAL REGRESSION NEURAL NETWORK BASED POS TAGGING FOR NEPALI TEXT

ArchitYajnik

Department of Mathematics, Sikkim Manipal University, Sikkim, India

## ABSTRACT

*This article presents Part of Speech tagging for Nepali text using General Regression Neural Network (GRNN). The corpus is divided into two parts viz. training and testing. The network is trained and validated on both training and testing data. It is observed that 96.13% words are correctly being tagged on training set whereas 74.38% words are tagged correctly on testing data set using GRNN. The result is compared with the traditional Viterbi algorithm based on Hidden Markov Model. Viterbi algorithm yields 97.2% and 40% classification accuracies on training and testing data sets respectively. GRNN based POS Tagger is more consistent than the traditional Viterbi decoding technique.*

## KEYWORDS

*General Regression Neural Networks, Viterbi algorithm, POS tagging*

## 1. INTRODUCTION

Artificial neural networks plays a vital role in various fields like medical imaging, image recognition is covered in [1, 2, 3] and since last one decade it becomes popular in the field of Computational linguistics also. Due to the computational complexities sometimes it is not preferred for the big data analysis. General Regression Neural Network which is based on Probabilistic neural networks is one type of supervised neural network is computationally less expensive as compared to standard algorithms viz. Back propagation, Radial basis function, support vector machine etc is exhibited in [4]. That is the reason GRNN is considered for the Past of speech Tagging experiment for Nepali text in this article.

Several statistical based methods have been implemented for POS tagging [5] as far as Indian languages are concern. Nepali is widely spoken languages in Sikkim and neighbouring countries like Nepal, Bhutan etc. The use of ANN architecture is seldom for tagging [6]. To develop a parser and Morphological analyser for the natural languages POS tagging plays a pivotal role.

This article presents a neural network architecture based on the Statistical learning theory described in [4]. This neural network is usually much faster to train than the traditional multilayer perceptron network. This article is divided into five sections. After this introduction, the second section presents the General Regression Neural Network from the Mathematical point of view

while the experimental set up of GRNN architecture is discussed in the third section. In the fourth section, the result analysis of POS Tagging using GRNN and Viterbi algorithm is presented for Nepali text followed by Conclusion in fifth section and references.

## 2. GENERAL REGRESSION NEURAL NETWORKS

The detailed information about Probabilistic and General Regression Neural Networks is available in [4]. GRNN can briefly be introduced for the training set,  $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ . To estimate the joint probability distribution for vectors  $\mathbf{x}$  and  $y$  say  $f_{\mathbf{x},y}(\mathbf{x}, y)$  and therefore  $f_{\mathbf{x}}(\mathbf{x})$ , we may use a nonparametric estimator known as the Parzen – Rosenblatt density estimator. Basic to the formulation of this estimator is a kernel, denoted by  $K(x)$ , which has properties similar to those associated with a probability density function:

Assuming that  $x_1, x_2, \dots, x_N$  are independent vectors and identically distributed (each of the random variables has the same probability distribution as the others), we may formally define the Parzen – Rosenblatt density estimate of  $f_{\mathbf{x}}(\mathbf{x})$  as

$$\hat{f}_{\mathbf{x}}(\mathbf{x}) = \frac{1}{Nh^{m_0}} \sum_{i=1}^N K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \text{ for } \mathbf{x} \in R^{m_0} \quad (1)$$

where the smoothing parameter  $h$  is a positive number called bandwidth or simply width;  $h$  controls the size of the kernel. Applying the same estimator on  $f_{\mathbf{x},y}(\mathbf{x}, y)$ , the approximated value for the given vector  $\mathbf{x}$  is given by

$$F(\mathbf{x}) = \hat{f}(\mathbf{x}) = \frac{\sum_{i=1}^N y_i K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)}{\sum_{i=1}^N K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)}$$

If we take Gaussian kernel i.e.  $K(\mathbf{x}) = e^{-\mathbf{x}^2}$ , we obtain,

$$\hat{f}(\mathbf{x}) = \frac{\sum_{i=1}^N y_i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^N \exp\left(-\frac{D_i^2}{2\sigma^2}\right)} \quad (2)$$

where  $D_i^2 = (\mathbf{x} - \mathbf{x}_i)^T (\mathbf{x} - \mathbf{x}_i)$  and  $\sigma$  is the standard deviation.  $\hat{f}(\mathbf{x})$  can be visualized as a weighted average of all observed values  $y_i$ , where each observed value is weighted exponentially according to its Euclidean distance from  $\mathbf{x}$ . The theory of General Regression Neural Networks discussed above is pertaining to only neuron in the output layer. The same technique can be

applied for the multiple neurons in the output layer also. Therefore the technique can be generalized as shown below:

Let  $w_{ij}$  be the target output corresponding to input training vector  $x_i$  and  $j^{\text{th}}$  output node out of the total  $p$ . Again let  $C_i$  be the centres chosen from the random vector  $x$ . Then

$$y_i = \frac{\sum_{i=1}^n w_{ij} h_i}{\sum_{i=1}^n h_i} \quad (3)$$

Here  $n$  be the number of patterns in the training set. The estimate  $y_j$  can be visualized as a weighted average of all the observed values,  $w_{ij}$ , where each observed value is weighted exponentially according to its Euclidean distance from input vector  $x$  and  $n$  is the number of patterns available in the input space.

$$\text{with } h_i = h_i(\sigma, C_i) = \exp\left(-\frac{D_i^2}{2\sigma^2}\right) \quad (4)$$

$$\text{where, } D_i^2 = (\mathbf{x} - \mathbf{C}_i)^T (\mathbf{x} - \mathbf{C}_i)$$

### 3. EXPERIMENTAL PROCEDURE AND RESULT

The survey of Part of Speech Tagging for Indian languages is covered by Antony P J (2011) in [7]. The details of the tags used for the experiment is available in [8, 9]. The total of 7873 Nepali words along with their corresponding text are collected. Out of which 5373 samples are used for training and the remaining 2500 samples for testing. The database is distributed in to  $n = 41$  tags. Network architecture consists of  $41 \times 3 = 123$  input neurons, 5373 hidden neurons which plays a role of centres  $C_i$  ( $i = 1, 2, \dots, 5373$ ) shown in (4) and 41 neurons in output layer.

#### 3.1 Feature Extraction and input neurons

Transition  $(T)_{n \times n}$  and Emission probability matrices  $(E)_{n \times m}$  are constructed for both the sets viz. training and testing. Transition matrix demonstrates the probability of occurrence of one tag (state) after another tag (state) hence becomes a square matrix  $41 \times 41$ . Whereas the emission matrix is the matrix of probability distribution of each Nepali word is allotted the respective tag hence it is of the size  $n \times m$  (number of Nepali words). In order to fetch the features for  $i^{\text{th}}$  word say  $x_i$ , the  $i^{\text{th}}$  row,  $i^{\text{th}}$  column of the transition matrix and  $i^{\text{th}}$  row of the emission matrix are combined hence becomes  $41 \times 3 = 123$  features for each word. Therefore the GRNN architecture consists of 123 input neurons.

#### 3.2 Hidden Neurons

All the patterns (or Nepali words) are used as a centre. Euclidean distance is calculated between patterns and centres. Training set consists of 5373 words hence the same number of hidden neurons are incorporated in GRNN architecture

### 3.3 Output Neurons

As there are 41 tags, 41 output neurons constitute the output layer of the network. For instance if the word belongs to NN (common noun) category which is the first tag of the tag set then the first neuron has a value 1 and all others are 0.

## 4. RESULT ANALYSIS

As the General Regression Neural Network follows supervised learning, the network is assigned 123 features as an input layer and 41 neurons as an output layer for each word. The network is trained using 5373 patterns (Nepali words) and corresponding tags. The network took 19 minutes to get trained. In the first phase, the same training set is validated, 4451 words out of 5373 are observed to be correct. In 715 words are tagged in the same group where they belong for example the word “गरिनेछ” has the actual tag “VBF (Finite Verb)” but it is assigned “VBX (Auxiliary Verb)” hence all together, the network has achieved 96.13% accuracy. In the second phase the network is tested on the words does not belong to the training set which contains 2500 Nepali words. The network has achieved 63.88% and 10.4% for correct identification and Group identification accuracy respectively, hence it becomes 74.28% total accuracy.

The same sets are tested using the traditional statistical technique Viterbi. The Viterbi decoding algorithm is applied for POS tagging for several natural languages[10]. The result obtained is depicted in table 1. The table emphasis that Viterbi algorithm gives very poor performance (40%) in identifying the words which do not belong to the training set by which the transition and emission matrices are constructed.

Table: 1 (Result using GRNN)

No	Technique	Validation set	Accuracy (%)	Group Identification accuracy (%)	Total Accuracy (%)
1	GRNN	Training set (5373)	82.84	13.29	96.13
2	GRNN	Testing set (2500)	63.88	10.4	74.28
3	Viterbi	Training set (5373)	93	4.2	97.2
4	Viterbi	Testing set (2500)	37	3	40

Percentage wise analysis is depicted in Fig. 1. The information shown in horizontal axis indicates that out of 320 total tags of NNP (Proper Noun), 200 are classified correctly and remaining 120 are confused with NN (Common Noun) as far as the first column is concern. Table 2 demonstrates the tags identified incorrectly and got confused with other tags. Proper Noun (NNP) is confused in 120 cases with the common noun (NN) because the probability of occurrence of NN after NNP is 0.62 while the reverse case has the probability of occurrence 0. That is the reason NNP is confused with NN but NN is never got confused with NNP even though both the tags belong to Noun group only as mentioned in table 1.2. The whole experiment is carried out in Java.



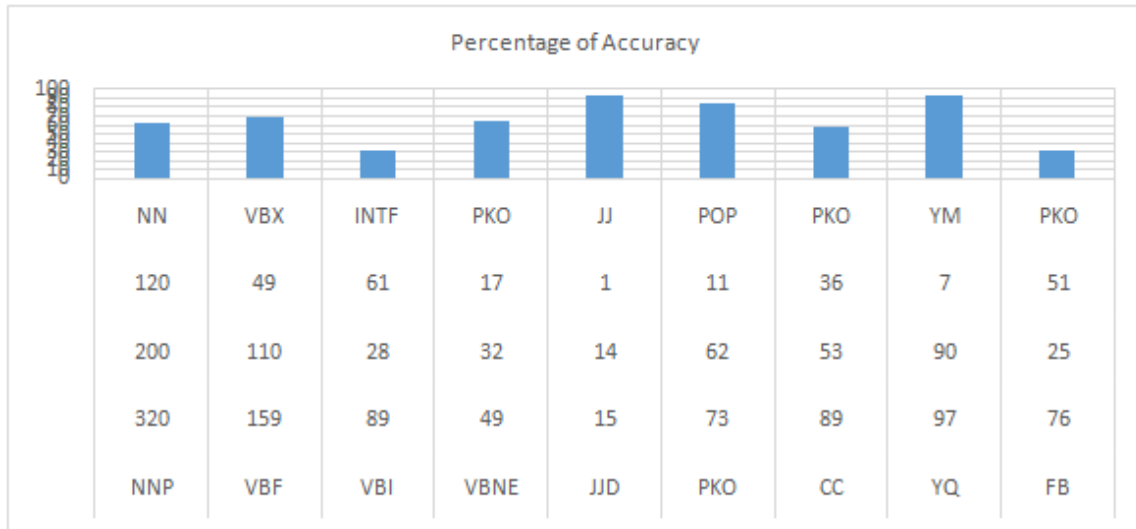


Figure. 1 Error Analysis using GRNN

Table 2 Confusion matrix

Tag	Frequency	Correct	Confusion	percentage	Group
NNP	320	200	NN	62.5	same
VBF	159	110	VBX	69.18	same
VBI	89	28	INTF	31.46	different
VBNE	49	32	PKO	65.31	different
JJD	15	14	JJ	93.33	same
PKO	73	62	POP	84.93	same
CC	89	53	PKO	59.55	different
YQ	97	90	YM	92.78	same
FB	76	25	PKO	32.89	different

## 5. CONCLUSIONS

In this article, GRNN based POS tagging approach is introduced for Nepali Text. Two techniques are employed viz. GRNN and Viterbi algorithm for this purpose. Section 4 demonstrates the outcome of the experiment on two data sets viz. training (5373 words) and testing (2500 words). Transition and emission probability matrices are constructed for both the techniques. Features are extracted from these matrices and used as an Input layer for GRNN as shown in section 3. Fully connected i.e. all the patterns (5373) are taken as centres GRNN architecture is trained using training set and outputs are validated on both training and testing sets. The result is compared with the traditional statistics based Viterbi algorithm. Table 1.1 shows that both the approaches yields satisfactory accuracy more than 96% as far as the training samples are concern but Viterbi fails completely (with 40% accuracy) while validated on testing set of 2500 patterns. On the other hand GRNN exhibits 74.28% accuracy. The confusion matrix (table 1.2) is generated on the output of GRNN on testing set. The accuracy may further be improved by collecting uniformly distributed data set.

## ACKNOWLEDGEMENTS

The author acknowledges Department of Science and Technology, Government of India for financial support vide Reference no SR/CSRI/28/2015 under Cognitive Science Research Initiative (CSRI) to carry out this work.

## REFERENCES

- [1] Richard O Duda and Peter E Hart, "Pattern Classification", 2006, Wiley-Interscience, New York, USA.
- [2] S. Rama Mohan, ArchiTajnik: "Gujarati Numeral Recognition Using Wavelets and Neural Network" Proceedings of Indian International Conference on Artificial Intelligence 2005, pp. 397-406.
- [3] ArchiTajnik, S. Rama Mohan, "Identification of Gujarati characters using wavelets and neural networks" Artificial Intelligence and Soft Computing 2006, ACTA Press, pp. 150-155.
- [4] Simon Haykin, "Neural Networks A Comprehensive Foundation" Second Edition, Prentice Hall International, Inc., New Jersey, 1999.
- [5] Prajadip Sinha et al. 2015. Enhancing the Performance of Part of Speech tagging of Nepali language through Hybrid approach, 5(5) International Journal of Emerging Technology and Advanced Engineering.
- [6] Tej Bahadur Shai et al. 2013. Support Vector Machines based Part of Speech Tagging for Nepali Text, Vol: 70-No. 24 International Journal of Computer Applications.
- [7] Antony P J et al. 2011. Parts of Speech Tagging for Indian Languages: A Literature Survey, International Journal of Computer Applications (0975-8887), 34(8).
- [8] <http://www.lancaster.ac.uk/staff/hardiea/nepali/postag.php>
- [9] <http://www.pan110n.net/english/Outputs%20Phase%202/CCs/Nepal/MPP/Papers/2008/Report%20on%20Nepali%20Computational%20Grammar.pdf>.
- [10] ArchiTajnik, "Part of Speech Tagging Using Statistical Approach for Nepali Text", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:11, No:1, 2017, pp. 76-79.

# SOCIAL NETWORK HATE SPEECH DETECTION FOR AMHARIC LANGUAGE

Zewdie Mossie<sup>1</sup> and Jenq-Haur Wang<sup>2</sup>

<sup>1</sup>Department of International Graduate Program in Electrical Engineering and  
Computer Science,

National Taipei University of Technology, Taipei, Taiwan

<sup>2</sup>Department of Computer Science and Information Engineering,  
National Taipei University of Technology, Taipei, Taiwan

## ABSTRACT

*The anonymity of social networks makes it attractive for hate speech to mask their criminal activities online posing a challenge to the world and in particular Ethiopia. With this ever-increasing volume of social media data, hate speech identification becomes a challenge in aggravating conflict between citizens of nations. The high rate of production, has become difficult to collect, store and analyze such big data using traditional detection methods. This paper proposed the application of apache spark in hate speech detection to reduce the challenges. Authors developed an apache spark based model to classify Amharic Facebook posts and comments into hate and not hate. Authors employed Random forest and Naïve Bayes for learning and Word2Vec and TF-IDF for feature selection. Tested by 10-fold cross-validation, the model based on word2vec embedding performed best with 79.83% accuracy. The proposed method achieve a promising result with unique feature of spark for big data.*

## KEYWORDS

*Amharic Hate speech detection, Social networks and spark, Amharic posts and comments*

## 1. INTRODUCTION

A major bottleneck for promoting use of computers and the Internet is that many languages lack the basic tools that would make it possible for people to access ICT in their own language. The status of language processing tools for European languages[2] states that only English, French and Spanish have sufficient basic tools. Thus the vast majority of the World's languages are still under-resourced in that they have few or no language processing tools and resources which particularly true for sub Saharan African languages. However, the evolution of the Internet and of social media texts, such as Twitter, YouTube and Facebook messages, has created many new opportunities for creating such tools, but also many new challenges [1]. Amharic is one of the sub-Saharan countries Ethiopian's working language which is written left-to-right in its own unique script which lacks capitalization and in total has 275 characters mainly consonant-vowel pairs. It is the second largest Semitic language in the world after Arabic and spoken by about 40% of the population as a first or second language [3] but current population estimated to 102 million. In

spite of its relatively large number of speakers, Amharic is still a language for which very few computational linguistic resources have been developed, and very little has been done in terms of making useful higher level Internet or computer based applications.

This paper focus only on hate speech detection from social media posts and comments. Recent advances in mobile computing and the Internet have resulted in an increase in use of social media to communicate, express opinions, interact with other, and to find and share information [4]. While social media provides an important avenue for communication to take place easily and efficiently, it also acts as a means of spreading hate speech online. Inherent characteristics of the Internet largely contribute to the misuse of social network to transmit and propagate hate speech.

Hate messages are prevalent and challenging in the Ethiopian online community as individuals spread hate messages hiding behind their screens. The government of Ethiopia oversee and monitor content in social network in a bid to govern hate speech through one time interruption of the internet service. Research conducted by Amnesty International and the Open Observatory of Network Interference (OONI) between June and October 2016 shows that access to WhatsApp and others was blocked, as well as at least 16 news outlets [6]. It is an open secret that the recent widespread hate speech and call for violence particularly targets persons of a particular group [5]. In this regard no work is done before and the first for Amharic language even though the work of [36] done from the social science perspective. It is therefore, of critical importance to monitor and identify instances of hate speech, as soon as possible to prevent their spread and possible unfolding into acts of violence or hate crimes and destroys the lives of individuals, families, communities and the country.

The proposed method used Word2Vec and TF-IDF for feature selection and Naïve Bayes and Random forest machine learning algorithms known for hate speech detection performance. The rest of this paper is organized as follows. Section 2 reviews related work on hate speech detection. The method and data preprocessing steps are described in detail in Section 3. Architectural design and experimentations are illustrated and discussed in Section 4. Finally, conclusion and future work in Section 5.

## **2. RELATED WORK**

### **2.1 Hate Speech on Social Media**

Online spaces are often exploited and misused to spread content that can be degrading, abusive, or otherwise harmful to people. Hateful speech has become a major problem for every kind of online platform where user-generated content appears from the comment sections of news websites to real-time chat sessions. Legal and academic literature generally defines hate speech as speech or any form of expression that expresses hatred against a person or group of people because of a characteristic they share, or a group to which they belong [7]. But, there is no consensus definition because of prevailing social norms, context, and individual and collective interpretation. A recent study define hate speech as speech which either promotes acts of violence or creates an environment of prejudice that may eventually result in actual violent acts against a group of people[8]. In the case of Ethiopia the use of hateful words with an intention to bring about hatred against a group of people based on their ethnicity, political attitude, religion and socio -economic are prevailing [36].

## 2.2 Social Media Definition of Hate Speech

- Hate speech is to incite violence or hate: The several definitions use slightly different terms to describe when hate speech occurs. The majority of the definitions point out that hate speech is to incite violence or hate towards a minority (Code of conduct, ILGA, YouTube and Twitter)
- Hate speech is to attack or diminish: Additionally, some other definitions state that hate speech is to use language that attacks or diminishes these groups (Facebook, YouTube, and Twitter).

After consulting those papers, authors use these dimensions of analysis to define what is hate speech in the scope of this paper.

## 2.3 Existing Techniques Used in Hate Speech Detection in Social Media

The study of hate speech detection has been growing only in the few last years. However, some studies have already been conducted in few languages. Papers focusing algorithms for hate speech detection, and also other studies focusing on related concepts, can give us insight about which features to use in this classification task. Therefore, authors allocate this specific section to describe the features already employed in previous works dividing into two categories: general features used in text mining and specific hate speech detection features.

**Dictionaries and lexicons:** The majority of the papers authors found try to adapt strategies already known in text mining to the specific problem of hate speech detection. The work categorize the features as the features commonly used in text mining which is dictionaries and lexicons. This approach consists in making a list of words that are searched and counted in the text. In the case of hate speech detection this has been conducted using content words such as insult and swear words, reaction words, and personal pronouns [24], number of disrespectful words in the text, with a dictionary that consists of words for English language including acronyms and abbreviations [26], label specific features which consisted in using frequently used forms of verbal abuse as well as widely used stereotypical words[27], Ortony lexicon was also used for negative affect detection ( list of words denoting a negative connotation and can be useful because not every rude comment necessarily contains bad language and can be equally harmful) [11].

**Bag-of-words(BOW):** Another model similar to dictionaries is the use of bag-of-words [9,10, 11]. In this case, a corpus is created based on the words that are in the training data, instead of a pre-defined set of words, as in the dictionaries. The disadvantages of this kind of approaches is that the word sequence is ignored, and also it's syntactic and semantic content. Therefore, it can lead to misclassification if the words are used in different contexts. To overcome this limitation n-grams were implemented. N-grams are one of the most used techniques in hate speech automatic detection and related tasks [11, 12, 13, 14, 15]. In a study character ngram features proved to be more predictive than to kennn-gram features, for the specific problem of abusive language detection [16].

**TF-IDF** was also used in this kind of classification problems. It is a measure of the importance of a word in a document within a corpus and increases in proportion to the number of times that a

word appears in the document. However, it is distinct from a bag of words, or n-gram, because the frequency of the term is off-set by the frequency of the word in the corpus, which compensates the fact that some words appear more frequently in general [17].

**Part-of-speech (POS)** approaches also make it possible to improve the importance of the context and detect the role of the word in the context of a sentence. These approaches consist in detecting the category of the word, for instance, personal pronoun (PRP), Verb non-third person singular present form (VBP), Adjectives (JJ), Determiners (DT), Verb base forms (VB). Part of speech has also been used in hate speech detection problem even though proved to cause confusion in the class's identification [14]. It was also used to detect sentences such as “send them home”, “get them out” or “should be hung” [18].

**Word Embedding:** Deep learning techniques are recently being used in text classification and sentiment analysis with high accuracy [28]. One of the approaches of this is word embedding which allows finding both semantic and syntactic relation of words, which permits the capturing of more refined attributes and contextual cues that are inherent in human language. Therefore, Word2Vec [19], an unsupervised word embedding-based approach to detect semantic and syntactic word relations was used. Word2Vec is a two-layer neural network that operates on a set of texts to initially establish a vocabulary based on the words included in such set more times than a user-defined threshold to eliminate noise. According to [19] 50-300 dimensions can model hundreds of millions of words with high accuracy. Possible methods to build the actual model are CBOW (i.e., Continuous bag of words), which uses context to predict a target word, and Skip-gram, which uses a word to predict a target context. Skip-gram works well with small amounts of training data and handles rare words or phrases well, while CBOW shows better accuracy for frequent words and is faster to train. Word embedding combined with Convolutional Neural Networks (CNN) show better performance [20, 28]. Authors [26] use a paragraph2vec approach to classify language on user comments as abusive or clean and also to predict the central word in the message. Alternatively, other authors propose comment embedding to solve this problem [27]. FastText is also being used [28] in a problem that sentences must be classified and not words. **Sentiment Analysis** bearing in mind that hate speech has a negative polarity, authors have been putting the sentiment as a feature for hate speech detection [15, 23, 24, 25, 31,].

## 2.4 Algorithms Commonly Used For Hate Speech Detection

Consulting different sources on algorithms of hate speech detection are the focus of this section, because authors aim to work in this specific topic. In the majority of the works the used language is English. However, there were some researched works done for languages Dutch [21] and Italian [22] to author's knowledge. The most common approach found in the work of [15] as a literature review consists in building a machine learning model for hate speech classification. It is found that the most common algorithms used are SVM, Random Forests, Decision Trees, logistic regression, Naïve Bayes and Deep learning respectively on the use of frequency by authors. The data classification is based on general hate speech, racism, sexism, religion, anti-Semitism, nationality, politics and socio-economics status respectively on the categorization use of frequency. Authors propose Random Forest and Naïve Bayes for their good performance.

### 3. PROPOSED METHODOLOGY AND DATA COLLECTION

Aiming at classifying the hate level across Facebook for Amharic language users, authors have built a corpus of comments retrieved from Facebook public pages of Ethiopian newspapers, individual politicians, activist, TV and radio broadcast and groups. These pages typically posts discussions spanning across a variety of political and religious topics. By doing so, authors could capture both casual conversations and politically hated posts and comments. Authors have employed a versatile Facebook crawler, which exploits the Graph API to retrieve the content of the comments from Facebook posts using Facepager. Facebook is selected to collect data from social media for the following reasons. Facebook is the most important platform for reaching out to online audiences, and especially the youth. Comparative studies have shown how in countries with limited Internet penetration, like Ethiopia, Facebook has become almost a synonym for the Internet, a platform through which users access information, services, and participate in online communications.

#### 3.1 Data Preparation and Annotation

Authors then preprocessed the posts and comments according to the following rules:

- Only kept comments that were in Amharic and all punctuations were removed by passing to the apache spark map function
- All null values are also removed with isNull attribute of apache spark DataFrame
- Checked to assure that no repetitions with the same text by passing to the map function using distinct attribute available on apache RDD and DataFrame
- Removed the HTML and different symbols in the same way using apache spark since authors focus only on texts
- All elongations were removed to the same fixed size character based on the nature of Amharic language and finally Trim text as final step

After all the above preprocessing authors consider the following three bases for future annotation:

(1) **Discourse analysis:**-places the text in its wider political, ethnicity, socio-economic and religious context in order to understand the currents of thought which illustrate and rationalize why it is to be considered hateful or not.

(2) **Content analysis:**-analyses the text deemed not hate and hate in order to pick out the key semantic components and targets of the speech. This can then be coded and quantitative techniques applied to draw wider patterns and trends.

(3) **Automated techniques:** - a relatively novel method of tracking hate speech that can be usefully employed to mine high volumes of text from different sources to search for keywords which are highly indicators of hate speech in an efficient manner which authors followed in labeling the collected data. After the initial cleaning authors got 25,890 posts and comments

available, however authors sampled to be 10, 000 due to the limitation in resources for the annotation task.

### 3.2 Annotation Instructions

Despite the differences between the previous studies that analyzed in the related work, the majority of the described works present instructions for the annotation task. Some authors point out that having vague annotation guidelines [8, 30] is a problem for hate speech detection due to the complexity of the task. In this work, authors prepared a complete set of annotation instructions in chart in order to better standardize the annotation procedure and to make clear all hate and not hate speech related category concepts. A set of instructions and examples that contain the indicators of the category was defined in figure 1. These are based on the definitions, rules and examples that presented already in the related work. The annotators were given the instructions as guidelines in the classification of the messages

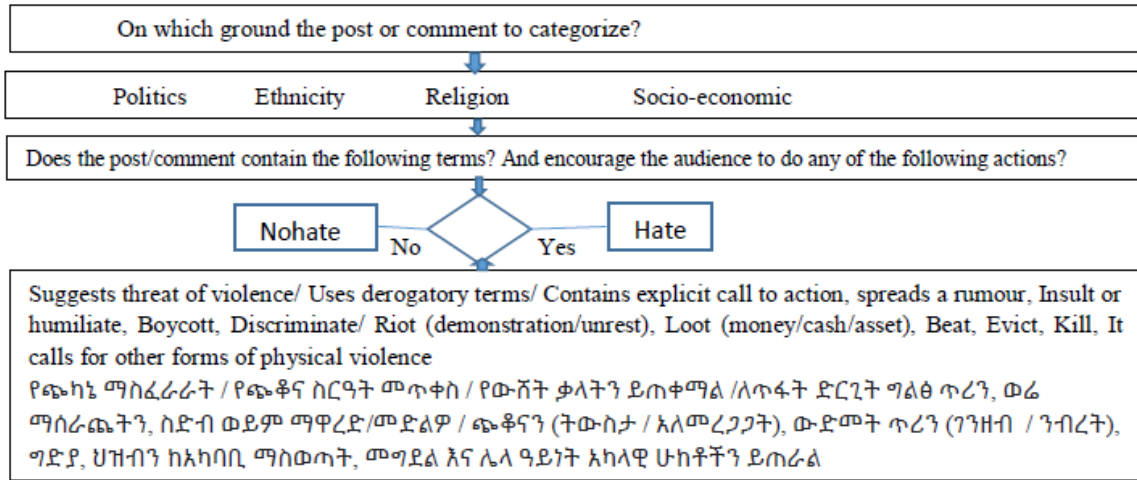


Figure 1: Hierarchical Structure of Dataset Annotation

In the work of [32] the training data was hand-coded and manually annotated and admits the potential for hard-to-trace bias in the hate speech categorization. The study concerned the detection of racism using a Naive Bayes classifier. The work established the definitional challenge of hate speech by showing annotators could agree only 33% of the time on texts purported to contain hate speech. Another considered the problem of detecting anti-Semitic comments in Yahoo news groups using support vector machines [29].

In this work first, authors consider a definition of hateful speech that could be practically useful to platform operators of social media and previous work definitions. Second, develop a general method of hierarchical annotation method shown in figure 1 for selected annotators of 3 PHD, 2 MSC students and 1 assistant professor from Amharic Language studies. The annotators were instructed to use the chart originates from the figure 1. In addition to the annotation rules the Kapa decision agreement based on the Cohen's kappa statistic which is an estimate of the population coefficient between  $0 \leq \kappa \leq 1$  [32] is also used. This work show how the values are interpreted? What does a specific kappa value mean?



Table 1: kappak values

Nominal	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa vlaue	0.0	0.20	0.40	0.60	0.80	1.0

Kappak agreement < 0 less than chance agreement, 0.01–0.20 Slight agreement, 0.21- 0.40 Fair agreement, 0.41-0.60 Moderate agreement, 0.61-0.80 Substantial agreement, 0.81-0.99, almost perfect agreement. Not everyone would agree which one is “good” agreement but as commonly cited scale is kappa value of 0.57 is in the “moderate” agreement range for better agreement. Remember that perfect agreement would equate to a kappa of 1, and chance agreement would equate to 0. Given that the majority of comments has been annotated by more than one annotator, authors have also computed the kappak inter-annotator agreement metric [33], which measures the level of agreement of different annotators on a task. In this case, considering 1,821 comments that received annotations from all the 6 annotators and obtain = 0.64 when discriminating over two classes and the work of [26] using number of disrespectful words in the text, with a dictionary that extracted from the annotated dataset and identified by the language experts. Then the dataset becomes larger than before which is 6, 120 to be used for this work.

Table 2: samples of kappak inter-annotator agreement result

Language	Amharic Comment Text and its English translation	class
Amharic	መፍትሄው ጎረቤትህን ለመጥፋት ነው።	Hate
English	The solution is to kill the neighboring Tigrian	
Amharic	ኦሮሞ የአማራ ጠላት ነው።	Hate
English	Oromo is an enemy of Amhara	
Amharic	አንድ አማራ ለሁሉም አማራ።	Nohate
English	One Amhara to All Amhara	

#### 4. ARCHITECTURAL DESIGN AND EXPERIMENTATION

An Apache Spark Standalone cluster was used for data preparation and developing models for machine learning classification which is suitable for big data processing like Facebook data. Spark ML pipeline is used in providing a set of tokenization mechanisms. In addition, Spark offers modules for feature selection and machine learning MLLIB library. Python programming language was used for both preparation of dataset and machine learning with RDD and DataFrame file format used as the back end for storing lazy operations which is ideal for large data. Spark designed to efficiently store RDD data while providing powerful MAP, Reduce and filter transformation operations and take actions for further process as shown in the figure 2.

The model was trained using 4,882 posts and used to correctly classify Facebook data according to the two classes mentioned above prepared based on the requirements of Naïve Bayes and Random forest algorithms. These classifier were selected based on previous work result in related work for English and other languages.

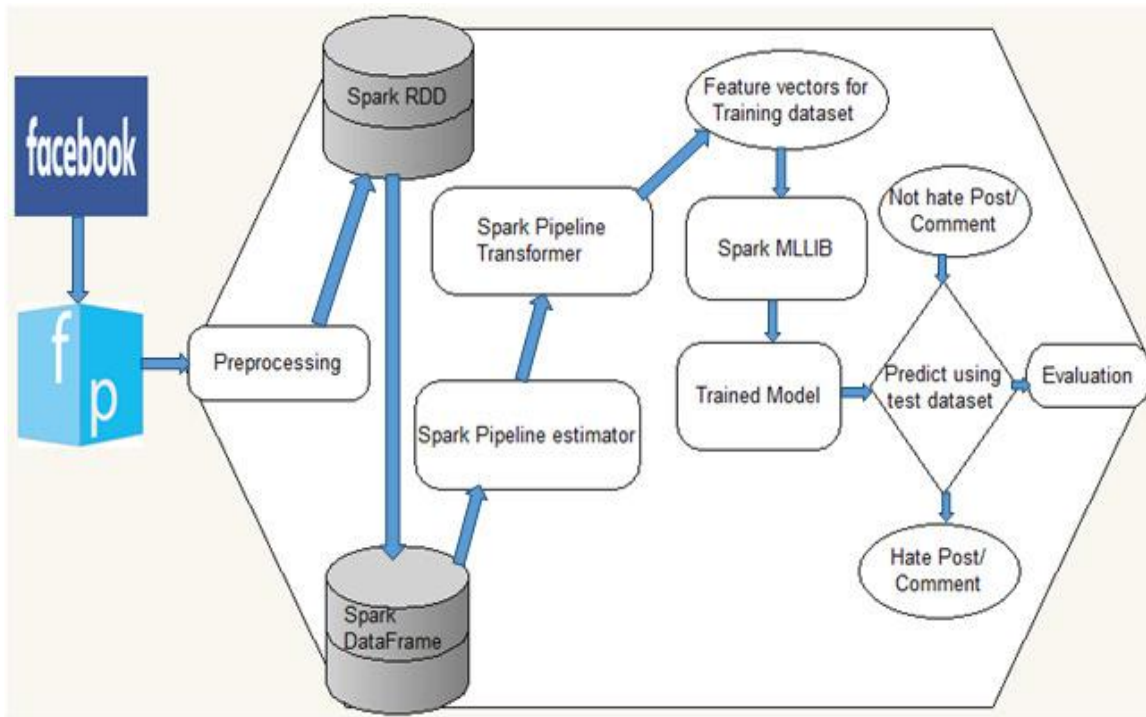


Figure 2: Architectural Design of Amharic hate speech detection

Table 2: Dataset information for the paper (new)

Training Dataset		
Nohate	Hate	Total
2,629	2,253	<b>4,882</b>
Test Dataset		
667	571	<b>1,238</b>
<b>3,296</b>	<b>2,824</b>	<b>6,120</b>

#### 4.1 Feature Selection

This involved selecting a subset of relevant features that would help in identifying hate and no hate posts and can be used in the modeling of the classification problems. Authors use Word2Vec as used in [11, 12, 13, 14] for such work and text classification [34]. TF-IDF [16] also used in text classification by different authors for feature selection in other tools. But authors propose to use both of them for Apache Spark feature selection and transformation API. The main feature of interest for this work is comments and posts sentiment of users towards hate speech in social media. The classification is supervised learning task because the objective is to use machine learning to automatically classify comments/posts into categories based on previously labelled comments and posts [11]. Author's contribution is preparation of new dataset, using tf-idf and word2vec as feature extraction, first in its kind, for the Amharic language hate speech detection proficient to big data on spark.

## 4.2 Model Design and Classification

To develop the model, Spark ML API (spark.ml) which provides ML pipelines (workflow) for creating, tuning, and evaluating of machine learning model was utilized. In Spark ML, a pipeline is defined as a sequence of stages, and each stage is either a Transformer or an Estimator. These stages are run in order, and the input DataFrame with spars vectors were transformed as it passes through each stage.

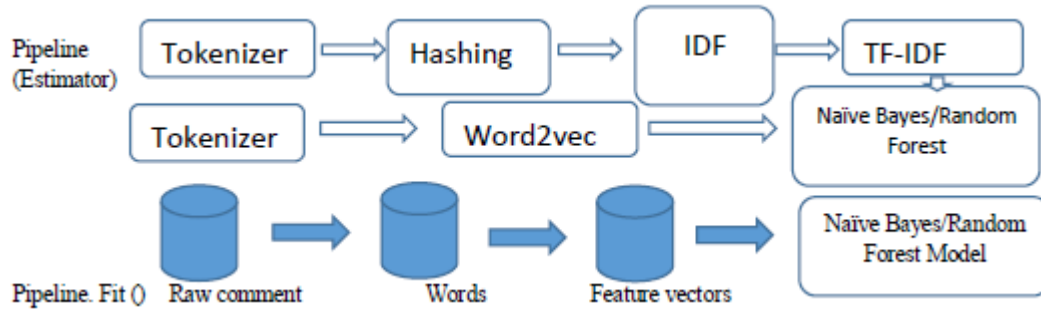


Figure 3: Spark ML pipeline for training (Adopted from Apache spark)

Annotated data were given to the pipeline to get features as feature vectors. The study split the dataset into two datasets, 80% (4882, comments) as training dataset and 20% (1238, comments) as testing dataset using the spark DataFrame random split function with the seed of 100. The training dataset was used to train model, and test dataset was used to evaluate the model performance.

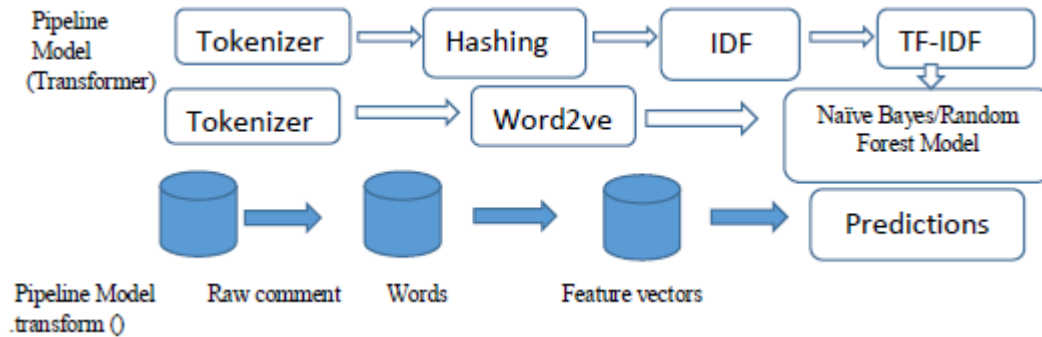


Figure 4: Spark ML pipeline for testing (Adopted from Apache spark)

## 4.3 Model Evaluation

For accuracy of the model, authors used cross validation using Spark evaluation tool namely Binary class Classification Evaluator within the spark ML. To evaluate the performance of the model classification in terms of quality or predictive effectiveness, different metrics appropriate for the work accuracy, ROC score and Area under curve F-measure (F1-score) were used as shown in table 3.

Table 3: Classification Performance result

Classifier Algorithm	Feature Model	Evaluation Metrics Result		
		Accuracy	ROC score	Area under PR
Naive Bayes	TF-IDF	0.73021	0.8053	0.7993
	<b>Word2Vec</b>	<b>0.7983</b>	<b>0.8305</b>	<b>0.8534</b>
Random Forest	TF-IDF	0.6355	0.6844	0.6966
	Word2Vec	0.6534	0.7097	0.7307

#### 4.4 Results and Analysis

Authors evaluated classification model by using the 10-fold cross-validation method, achieving an average result as presented in table 3. It was evident that the Naïve Bayes classifier with word2Vec feature model outperform to classify hate and Nohate speech 0.7983, 0.8305 and 0.8534 accuracy, ROC score and area under Precision and Recall respectively with Facebook social network for Amharic language posts and comments. The Naïve Bayes also achieve better result for TF-IDF feature model with 0.73021, .08053 and 0.7993 for accuracy, ROC score and area under precision and recall respectively. The Random Forest with word2vec feature is better than TF-IDF with the result 0.6534, 0.7097 and 0.7307 accuracy, ROC score and area under precision and recall respectively. This is followed by TF-IDF with 0.6355, 0.6844 and 0.6996 respectively. Even though may not be appropriate to compare the result with different experimental setups authors got the state of the art result found in other languages with unique feature of scalability for big data.

The following two charts shows sample of the hate speech classification performance using ROC score area.

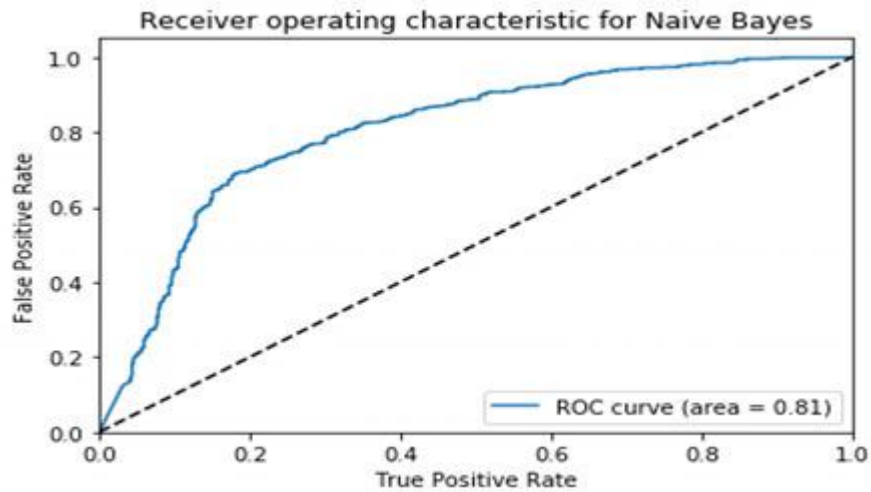


Figure 5: ROC for Naïve Bayes with TF-IDF

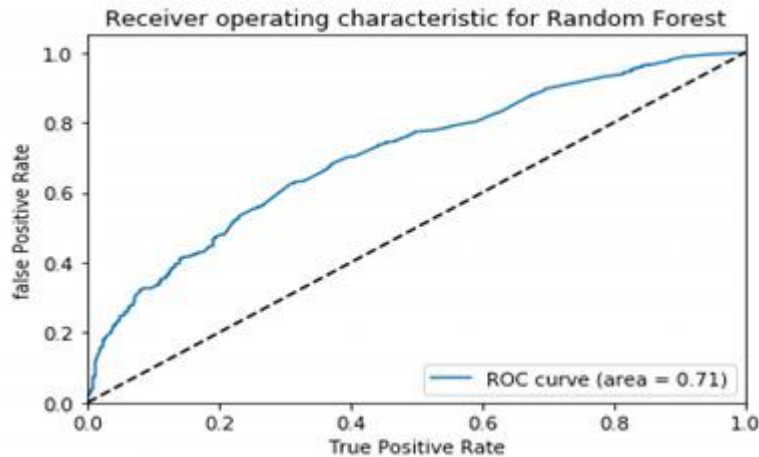


Figure 6: ROC for Random Forest with Word2Vec

## 5. CONCLUSIONS AND FUTURE WORK

The study developed a model for Amharic text hate speech detection that analyzes posts and comments to identify hate speech using spark machine learning techniques. To conduct the experiments, thousands of Amharic post and comments on suspected social network pages of organizations and individual person's public pages are crawled as dataset. First preprocessed according to the requirement of the language and human annotators selected to label the comment in to hate or not hate. Here after, features are pipelined to word2vec neural network tool and TF-IDF in apache spark environment so that feature vectors are obtained.

The classification algorithms were implemented in Apache Spark local cluster using the Apache Spark's Machine Learning library. The model developed using Naïve Bayes and Random forest utilizing a dataset of 6,120 Amharic posts and comments out of this 4,882 to train the model and 1,238 for testing after passing different steps as stated in the experiment section. The model was tested to classify whether the post and comments are hate or not and able to detect and classify in an accuracy of 79.83 % and 65.34% for Naïve Bayes with word2vec feature vector and Random Forest with TF-IDF feature modeling approach respectively. The workshow that word2vec feature model is better in maintaining the semantics of the posts and comments as proved in other works. The result are promising for such work in social network big data which can be extended to compute large volumes data since the work used the distributed platform of apache spark.

Even if the results are promising for hate detection, our research is far from perfect. A lot of work ahead of us to work on technical improvements that can be made for the language interms of: (1) expand the dataset that would reduce the risk of overfitting and improve the statistical significance of the results (2)analyzing the different aspect of the category of hate, either hate with politics, ethnicity, religion and socio-economy (3)utilize the information provided by Facebook so that, classification can be improved by expanding the feature space with profile information, list of followers and geolocation etc. (4) crawl other sources to improve the feature space for such under resourced language for computational purpose by adding synonyms from other sources such as Twitter, forums and other homepages.

Finally, the proposed methods could be applied in different domains where the posts about the anticipation to get service and buy product by the review of the service after serving or buying it for this particular language showing the sentiment of costumers as positive or negative can be explored.

## REFERENCES

- [1] Bjorn Gambäck and Utpal Kumar Sikdar, Named Entity Recognition for Amharic Using Deep Learning. IST-Africa 2017 Conference Proceedings Paul Cunningham and Miriam Cunningham (Eds) IIMC International Information Management Corporation, ISBN: 978-1-905824-57-1, 2017
- [2] META-NET White Paper Series, Retrieved from Multilingual Europe Technology Alliance: <http://www.meta-net.eu/whitepapers/overview> , (2018, January Tuesday)
- [3] Grover Hudson, Linguistic analysis of the 1994 Ethiopian census, *Northeast African Studies*, 6(3):89–107, 1999
- [4] Raphael Cohen-Almagor. Internet History, *International Journal of Techno ethics*, Vol. 2, No. 2, pp. 45-64, 2011
- [5] Waseem & Hovy, Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of NAACL-HLT*, 2016, pages 88–93
- [6] White Paper series, Retrieved from Amnesty International, Social media and Internet: <https://www.amnesty.org/en/latest/news/2016/12/ethiopia-government-blocking-of-websites-during-protests-widespread-systematic-and-illegal/>, 2016, 2018.
- [7] Saleem, Haji Mohammad, Kelly P. Dillon, Susan Benesch, and Derek Ruths, A web of hate: Tackling hateful speech in online social spaces. *ArXiv preprint arXiv: 1709.10159*, 2017.
- [8] Fortuna, Paula Cristina Teixeira, Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes, 2017.
- [9] Kwok, Irene, and Yuzhou Wang, Locate the Hate: Detecting Tweets against Blacks, In *AAAI*. 2013
- [10] Del Vigna<sup>12</sup>, Fabio, Andrea Cimino<sup>23</sup>, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi, Hate me, hate me not: Hate speech detection on Facebook, 2017.
- [11] Silva, Leandro Araújo, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber, Analyzing the Targets of Hate in Online Social Media, In *ICWSM*, pp. 687-690, 2016.
- [12] Waseem, Zeerak, and Dirk Hovy, Hateful symbols or hateful people? Predictive features for hate speech detection on twitter, In *Proceedings of the NAACL student research workshop*, pp. 88-93, 2016.
- [13] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang, Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145-153. International World Wide Web Conferences Steering Committee, 2016.
- [14] Davidson, Thomas, Dana Warmusley, Michael Macy, and Ingmar Weber, Automated hate speech detection and the problem of offensive language, *arXiv preprint arXiv: 1703.04009*, 2017.

- [15] Yashar Mehdad and Joel Tetreault, Do characters abuse more than words? In Proceedings of the SIGdial 2016 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 299–303, 2016.
- [16] Keith Cortis and Siegfried Handschuh, Analysis of cyberbullying tweets in trending world events. In Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business, page 7. ACM, 2015.
- [17] Agarwal, Swati, and Ashish Sureka, Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website, arXiv preprint arXiv: 1701.04931, 2017.
- [18] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv: 1301.3781, 2013.
- [19] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, Distributed representations of words and phrases and their compositionality, In Advances in neural information processing systems, pp. 3111-3119, 2013.
- [20] Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki, Measuring the reliability of hate speech annotations: The case of the European refugee crisis, 2017.
- [21] Stephan Tulkens, Lisa Hilde, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans, A dictionary-based approach to racism detection in Dutch social media, 2016.
- [22] Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi, Hate me, hate me not: Hate speech detection on Facebook. In Proceedings of the First Italian Conference on Cybersecurity, pages 86–95, 2017.
- [23] Liu, Shuhua, and Thomas Forss, Combining n-gram based similarity analysis with sentiment analysis in web content classification, In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management-Volume 1, pp. 530-537. SCITEPRESS-Science and Technology Publications, Lda, 2014.
- [24] Liu, Shuhua, and Thomas Forss, New classification models for detecting Hate and Violence web content, In Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on, vol. 1, pp. 487-495. IEEE, 2015.
- [25] Maloba, Wilson Jeffrey, Use of regular expressions for multi-lingual detection of hate speech in Kenya, PhD diss., iLabAfrica, 2014.
- [26] Njagi Dennis Gitari, Zhang Zuping, Hanyurwim fura Damien, and Jun Long. A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering, 10(4):215–230, 2015.
- [27] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati, Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web, pages 29–ACM2, 2015.
- [28] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma, Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

- [27] Maral Dadvar, Franciska de Jong, Roeland Ordelman, and Dolf Trieschnigg, Improved cyberbullying detection using gender information. In Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop, pages 23–25, University of Ghent, 2012.
- [28] Shuhan Yuan, Xintao Wu, and Yang Xiang, A two phase deep learning model for identifying discrimination from tweets, In International Conference on Extending Database Technology, pages 696–697, 2016.
- [29] Kwok and Wang, Detecting Tweets against Blacks. Proceedings of the Twenty-Seventh AAI Conference on Artificial Intelligence, 2013
- [30] William Warner and Julia Hirschberg, Detecting Hate Speech on the World Wide Web. Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pages 19–26, Association for Computational Linguistics, Montreal, Canada, 2012
- [31] Anna Schmidt and Michael Wiegand, A survey on hate speech detection using natural language processing. Social NLP 2017, page 1, 2017.
- [32] Anthony J. Viera, Understanding Inter observer Agreement: The Kappa Statistic, From the Robert Wood Johnson Clinical Scholars Program, University of North Carolina, 2005
- [33] Kilem L Gwet. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.
- [34] Joseph Lilleberg et al, Support Vector Machines and Word2vec for Text Classification with Semantic Features. Proc. 2015 IEEE 14th Int'l Coni. On Cognitive Informatics & Cognitive Computing IEdsJ, 2015.
- [35] United Nations Educational, Scientific and Cultural Organization, Countering Online Hate Speech. Published in 2015 by the United Nations Educational, Scientific and Cultural Organization 7, place de Fontenot, 75352 Paris 07 SP, France
- [36] Iginio Gagliardone, Alisha Patel and Matti Pohjonen, Mapping and Analyzing Hate Speech Online: Opportunities and Challenges for Ethiopia, 2014



**AppendixA****Experiment setup**

Apache environment setup	
Apache Spark-2.2.0	Ubuntu 16.4 virtual machine Standalone clustering mode 1 master node 2 worker node
Feature selection Algorithms	
Name of Algorithm	Parameter set up
Word2Vec	choice of training model 0: Skip gram model dimension of vectors=3 Window size=5 Minimum count=0
TF-IDF	Default
Machine learning Algorithms	
Name of Algorithm	Parameter set up
Naïve Bayes	Model type =multinomial Smoothing =1
Name of Algorithm	Parameter set up
Random Forest	NumTree=200 MaxDepth=3 Seed=2
10-fold cross-validation	
Name of Algorithm	Parameter set up
Naïve Bayes	Smoothing =1 Numfolds=10
Random Forest	NumTree=[50,100,200] MaxDepth=[3,4,5] Numfolds=10

*INTENTIONAL BLANK*

# SOCIAL MEDIA ANALYTICS FOR SENTIMENT ANALYSIS AND EVENT DETECTION IN SMART CITIES

Aysha Al Nuaimi, Aysha Al Shamsi and Amna Al Shamsi, Elarbi Badidi

College of Information Technology, United Arab Emirates University, Al-Ain,  
United Arab Emirates

## ABSTRACT

*Smart cities utilize Internet of Things (IoT) devices and sensors to enhance the quality of the city services including energy, transportation, health, and much more. They generate massive volumes of structured and unstructured data on a daily basis. Also, social networks, such as Twitter, Facebook, and Google+, are becoming a new source of real-time information in smart cities. Social network users are acting as social sensors. These datasets so large and complex are difficult to manage with conventional data management tools and methods. To become valuable, this massive amount of data, known as 'big data,' needs to be processed and comprehended to hold the promise of supporting a broad range of urban and smart cities functions, including among others transportation, water, and energy consumption, pollution surveillance, and smart city governance. In this work, we investigate how social media analytics help to analyze smart city data collected from various social media sources, such as Twitter and Facebook, to detect various events taking place in a smart city and identify the importance of events and concerns of citizens regarding some events. A case scenario analyses the opinions of users concerning the traffic in three largest cities in the UAE*

## KEYWORDS

*Internet of things, Urban data streams, Stream processing, Big data, Analytics*

## 1. INTRODUCTION

Modern cities use digital technologies to reduce costs, balance budgets, enhance the efficiency of various city systems, optimize city management, improve the quality of services delivered to citizens, create new facilities for the public, reduce energy consumption and thus offer a better quality of urban life. These technologies create new opportunities for cities to make themselves smarter through innovative planning and information-based management and operation. As the amount of data, in structured and unstructured formats, is so huge, new approaches to data management are needed [1]. Besides, social networks (such as Twitter, Facebook, and Google+) are becoming a new source of real-time information in smart cities. Social network users are regarded as social sensors. To become valuable, this massive amount of data, known as 'big data,' needs to be processed and comprehended.

Big data analytics is a recent technology that has an immense potential to permit comprehending city data and, hence, enhancing smart city services. Effective management and analysis of big data is a fundamental component to achieve the goals of the smart city. These goals include tackling the problems, reducing resources consumption and costs, engaging actively with citizens, and making informed decisions that will result in enhancing the environment and improving economic outcomes leading to improved quality of urban life. For the big data scientist, there is opportunity amongst this vast amount and array of data. The analysis of big data has the potential to improve the lives of citizens and reduce costs by uncovering associations and by understanding the trends and patterns in the data.

In this work, we propose to collect and analyze social media data concerning the main cities in the UAE. An increasing number of events is taking place every year in these cities. Therefore, it is paramount to analyze the conversations of the citizens with regards to these events and other concerns. First, we focus on data retrieved from Twitter given the small size of the tweets. Twitter is an exciting source of information for real-time event detection and sentiment analysis. Twitter API is used to get up-to-date tweets. As social media analytics uses several analysis and modeling techniques, we focus primarily on techniques such as sentiment and trend analysis that support the data understanding phase. This quest will contribute to a better understanding of the needs and concerns of the public so that event organizers and municipal governments take appropriate action to address these concerns.

The remainder of the paper is organized as follows. Section 2 provides background information on the benefits of big data analytics in smart cities and the application of social media analytics for events' detection. Section 3 describes the methodology followed to analyze social media data streams. Section 4 describes a case scenario in which tweets concerning the UAE are analyzed by applying sentiment analysis. Finally, Section 5 concludes the paper and highlights future work.

## **2. BACKGROUND AND LITERATURE REVIEW**

### **2.1 Big data analytics applications**

Using advanced analytics techniques such as data mining, machine learning, statistical learning, text analytics, predictive analytics, and visualization tools, city stakeholders and local governments would be able to analyze previously inaccessible or unusable data to gain new insights resulting in significantly faster and informed decisions. These techniques can speed up the analytical investigation, leading to insights from both traditional and non-traditional data sources. Ideally, city and local government would use data analytics to monitor public utilities, alleviate traffic congestion, assess and anticipate crime, follow education trends, and carefully watch public resources.

Another smart city application where big data analytics will play a vital role is crowd control. Indeed, the cities are more and more crowded, and many events are organized with as many people attending the events. As a result, municipalities should provide many services to the participants such as safety, mobilizing police and emergency responders, basic needs such as food and beverages, and many more. Crowds of people mean that massive amounts of data are generated. Big data analytics can be used to predict the movement of the masses to avoid jostling or other severe disasters.

During emergencies, cities equipped with sensors can take advantage of collected data to make informed decisions. Using social media analytics during emergency response is of particular interest. Social media networks offer data streams, which can be used to collect near real-time information concerning an emergency. Though people post on social networks many unrelated messages, any emergency information can be valuable to emergency response teams and can help them to get a good picture of the situation, permitting a more effective and faster response that can reduce overall loss and damage [2].

## **2.2 Social media analytics**

Social network sites, such as Facebook, Twitter, and Google+, have become so popular that they represent a new source of real-time information concerning various topics and events. The availability of social networks sites on numerous devices ranging from personal computers to tablets and smartphones contributed to their popularity [3][4].

People typically use social media networks to post small messages that allow them to express their opinions on a variety of topics or to report events occurring in their vicinity. These networks allow their users to have an identity, build small online communities, find other users with similar interests, and find content published by other users [5]. Shared messages on these networks are called Status Update Messages (SUM). In addition to the text of the message, a SUM contains metadata information such as the name of the user, timestamp, geographic coordinates (latitude and longitude), hashtags, and links to other resources.

The SUMs originating from users in a specific geographical area, such as a city, or discussing a topic or raising a concern can offer valuable information on an issue or event that can help decision-makers make informed decisions. As a result, social network users are considered as social sensors [6][7], and SUMs as sensor information [8].

Social media analysis is about collecting data from social media websites and blogs and processing that data as structured information that can lead decision-makers to make information-driven decisions. The customer is at the center of these decisions. The most common use of social media analytics is to leverage customers' opinions about products and services to support customer service and marketing activities.

## **2.3 Detection of events from social media analytics**

In recent years, social networks have become a new source of information that municipal authorities can use to detect events, such as traffic jams, and get reports on incidents and natural disasters (fires, storms, tremors etc.) in their neighborhood. An event is a real-world occurrence that happens at a given moment and space [3] [9]. In particular, regarding traffic-related incidents, people often share using SUM information about the current traffic situation around them while driving.

However, event detection from the analysis of social networks' short-messages is more challenging compared with event detection from conventional media, such as emails and blogs, where texts are well-formatted [4]. SUMs are unstructured and irregular texts, which might contain simple or abbreviated words, grammatical errors or misspellings [3]. They are usually very brief, which makes them an incomplete source of information [4]. Moreover, SUMs

contain a massive amount of meaningless or not useful information [10], which require filtering. To extract meaningful information from social media networks, it is necessary to use text mining techniques that use methods in the fields of Natural Language Processing (NLP), machine learning, and data mining. [11] [12].

Concerning current approaches for using social media to obtain useful information for event detection, there is a need to distinguish between large-scale events and small-scale events. Large-scale events such as the election of a president, earthquakes or the tsunami are generally characterized by a considerable number of SUMs and a wider temporal and geographic coverage. Instead, small events such as traffic jams, car accidents, fires or local shows are characterized by a small number of associated SUMs, limited geographic and temporal coverage [21]. As a result, small-scale event detection is not a trivial task because of the smaller number of SUMs related to these kinds of events. Several research efforts studied event detection from social networks data. Many of them deal with large-scale event detection [8][13][14][15][16] and only a few of these works investigated small-scale events [17][18][19][20].

### 3. METHODOLOGY

A three-phase process, which includes data capture, data understanding, and presentation is used in this work (see Figure 1).

*Data Capture.* This phase helps identify messages on social media networks about the city's activities, events, and concerns. This process is achieved by collecting massive amounts of pertinent data from social media networks using primarily APIs, provided by these sources, or through crawling. The main popular social media platforms are Twitter, Facebook, Google+, LinkedIn, and YouTube. Some preprocessing steps may be performed to prepare a dataset for the data understanding phase. They typically include modeling and data representation, syntactic and semantic operations, linking data originating from different sources, and feature extraction.

*Data understanding.* In this phase, the meaning of collected data in the previous phase is assessed and metrics useful for decision-making are generated. As collected data might originate from multiple sources, it might contain a significant amount of noise that need to be cleaned before conducting meaningful analyses. This cleaning function may use simple, rule-based text classifiers or more sophisticated classifiers trained on labeled data. To assess the meaning of cleaned data, various techniques such as text mining, natural language processing, social network analysis, and statistical methods can be used. This phase can provide information about user opinions concerning an event or a service. Its results significantly impact the information and metrics in the presentation phase as well as the future decisions and actions a smart city might take.

*Presentation.* In this last phase, the results of different analyzes are summarized, evaluated and presented to users in an easy-to-understand format. The visual dashboard, which aggregates and displays data from multiple sources, is a commonly used interface design.

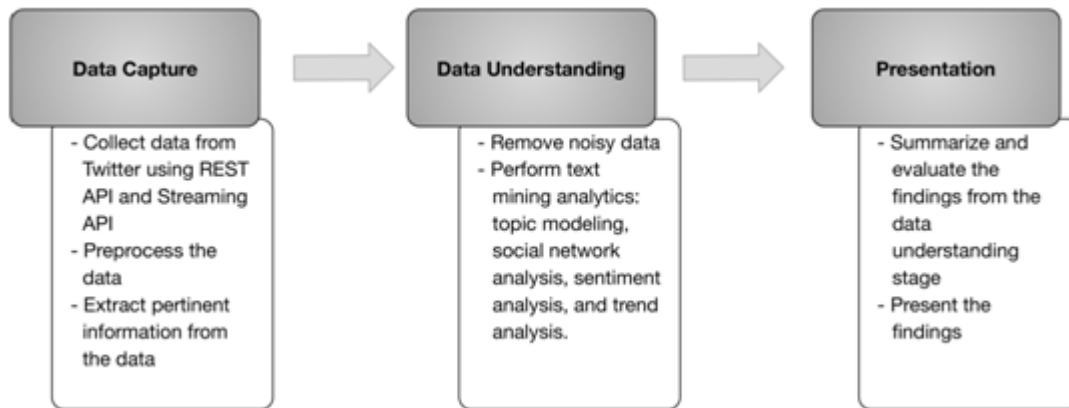


Figure 1. Overview of the proposed method

## 4. CASE SCENARIO

In this scenario, we implement a real-time application to get the latest Twitter feeds concerning the #Dubai, #AbuDhabi, and #Sharjah hashtags and process them by subjecting them to sentiment analysis.

Data streams (tweets) from Twitter have been recognized as a valuable data source for many smart cities in areas such as law enforcement, tourism, and politics (e.g., US presidential election). The Twitter Streaming API permits extracting datasets that are then used to perform sentiment analysis.

The following workflow describes the different steps we used for the analysis of the tweets:

1. Authenticate and connect to Twitter using Twitter API
2. Use search REST API to gather tweets from an Input file using keywords, hashtags, and Twitter account.
3. Store collected tweets in a CSV file
4. Preprocess the gathered tweets by removing duplicates
5. Use TextBlob to obtain sentiment for the unique tweets. TextBlob is a Python library for processing textual data. The result of a sentiment analysis task is the percentage of positive, negative, and neutral opinions.
6. Compute total, positive, negative and neutral tweets over a number of days.
7. Visualize results

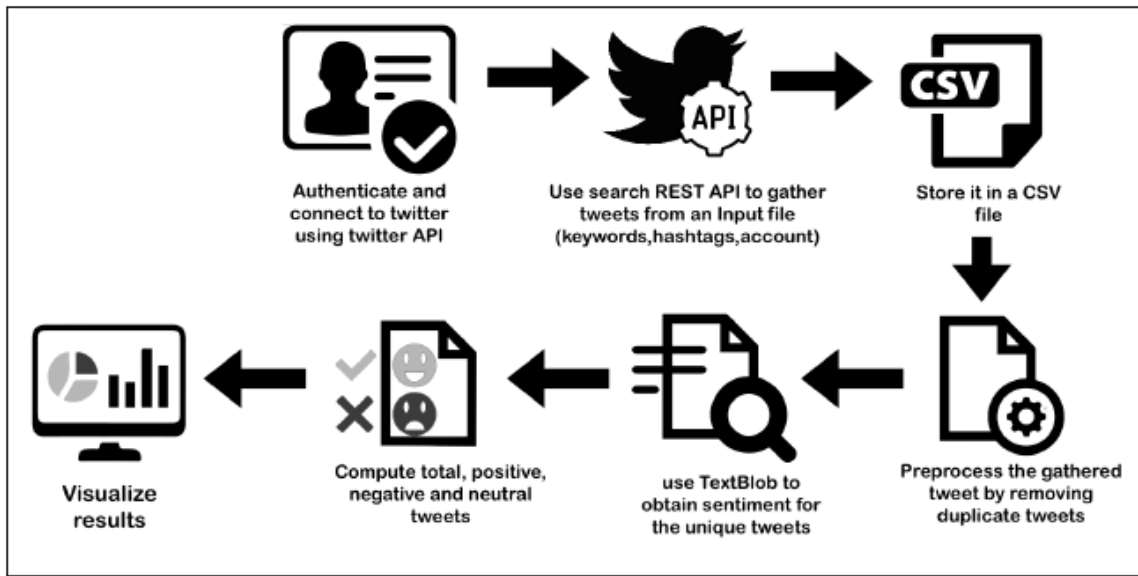


Figure 2.A workflow for the sentiment analysis of tweets

Figure 3 depicts the results obtained by analysing the opinions of the users regarding the traffic in the UAE main cities: Dubai, Abu Dhabi, and Sharjah over a week period from March 7, 2018, to March 12, 2018.

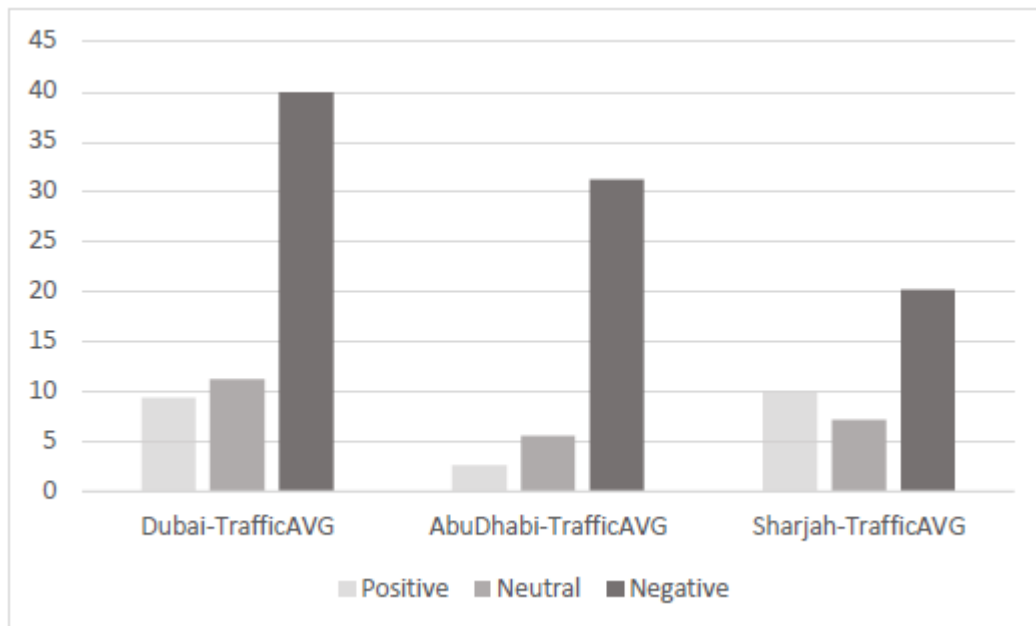


Figure 3. Opinions of the users regarding the traffic in the UAE main cities



The results show that a significant proportion of users have a negative opinion on the traffic in the three large UAE cities with more negative views on the traffic in Dubai. The percentage of users with a positive opinion is relatively low compared with negative opinions in the three cities. However, for Sharjah, that percentage is a little bit higher than the percentage of neutral opinions, which is not the case for Dubai and Abu Dhabi. These results indicate that users are suffering from the traffic conditions in the three cities and that municipal services should work hard to find solutions to the traffic problem, which is one of the challenges that most megacities are facing.

## 5. CONCLUSION

Modern cities are more and more relying on the usage of the Internet of Things (IoT) devices and sensors to sense various parameters such as temperature humidity, water leaks, sunlight, and air pressure. These devices and sensors generate massive volumes of data. Also, social networks are becoming a new source of real-time information in smart cities. Social network users are acting as social sensors. In this work, we described how social media analytics could help analysing urban data streams collected from popular social media sources, such as Twitter and Facebook, to detect events taking place in a smart city and identify the concerns of citizens regarding some events or issues. We analyse in a case scenario the sentiments of users concerning the traffic in three largest cities in the UAE. The results show how frustrated are the users with the traffic in the cities of Dubai and Abu Dhabi that their municipalities need to work hard to alleviate this issue.

## REFERENCES

- [1] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, (2016) "The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748–758.
- [2] Tim L. M. van Kasteren, Birte Ulrich, Vignesh Srinivasan, Maria E. Niessen, (2014) "Analyzing Tweets to Aid Situational Awareness," *Advances in Information Retrieval*, Vol. 8416, Lecture Notes in Computer Science, pp 700-705.
- [3] F. Atefeh and W. Khreich, (2015) "A survey of techniques for event detection in Twitter," *Comput. Intell.*, vol. 31, no. 1, pp. 132–164.
- [4] P. Ruchi and K. Kamalakar, (2013) "ET: Events from tweets," in *Proc. 22nd Int. Conf. World Wide Web Comput.*, Rio de Janeiro, Brazil, pp. 613–620.
- [5] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, (2007) "Measurement and analysis of online social networks," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, San Diego, CA, USA, pp. 29–42.
- [6] G. Anastasi et al., (2013) "Urban and social sensing for sustainable mobility in smart cities," in *Proc. IFIP/IEEE Int. Conf. Sustainable Internet ICT Sustainability*, Palermo, Italy, pp. 1–4.
- [7] A. Rosi et al., (2011) "Social sensors and pervasive services: Approaches and perspectives," in *Proc. IEEE Int. Conf. PERCOM Workshops*, Seattle, WA, USA, pp. 525–530.
- [8] T. Sakaki, M. Okazaki, and Y. Matsuo, (2013) "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 919–931.

- [9] J. Allan, (2002) “Topic Detection and Tracking: Event-Based Information Organization”. Norwell, MA, USA: Kluwer.
- [10] J. Hurlock and M. L. Wilson, (2011) “Searching Twitter: Separating the tweet from the chaff,” in Proc. 5th AAAI ICWSM, Barcelona, Spain, pp. 161–168.
- [11] S. Weiss, N. Indurkha, T. Zhang, and F. Damerau, (2004) “Text Mining: Predictive Methods for Analyzing Unstructured Information,” Berlin, Germany: Springer-Verlag.
- [12] Hotho, A. Nürnberger, and G. Paaß, (2005) “A brief survey of text mining,” LDV Forum-GLDV J. Comput. Linguistics Lang. Technol., vol. 20, no. 1, pp. 19–62.
- [13] M. Krstajic, C. Rohrdantz, M. Hund, and A. Weiler, (2012) “Getting there first: Real-time detection of real-world incidents on Twitter” in Proc. 2nd IEEE Work Interactive Vis. Text Anal.—Task-Driven Anal. Soc. Media IEEE VisWeek,” Seattle, WA, USA.
- [14] C. Chew and G. Eysenbach, (2010) “Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak,” PLoS ONE, vol. 5, no. 11, pp. 1–13.
- [15] B. De Longueville, R. S. Smith, and G. Luraschi, (2009) “OMG, from here, I can see the flames!: A use case of mining location based social networks to acquire spatiotemporal data on forest fires,” in Proc. Int. Work. LBSN, Seattle, WA, USA, pp. 73–80.
- [16] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, (2012) “Using social media to enhance emergency situation awareness,” IEEE Intell. Syst., vol. 27, no. 6, pp. 52–59.
- [17] T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri, and K. Nawa, (2012) “Real-time event extraction for driving information from social sensors,” in Proc. IEEE Int. Conf. CYBER, Bangkok, Thailand, pp. 221–226.
- [18] P. Agarwal, R. Vaithyanathan, S. Sharma, and G. Shro, (2012) “Catching the long-tail: Extracting local news events from Twitter,” in Proc. 6th AAAI ICWSM, Dublin, Ireland, Jun. pp. 379–382.
- [19] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, (2012) “Twitcident: fighting fire with information from social web streams,” in Proc. ACM 21st Int. Conf. Comp. WWW, Lyon, France, pp. 305–308.
- [20] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, (2012) “TEDAS: A Twitter- based event detection and analysis system,” in Proc. 28th IEEE ICDE, Washington, DC, USA, pp. 1273–1276.
- [21] A. Schulz, P. Ristoski, and H. Paulheim, ( 2013) “I see a car crash: Real-time detection of small scale incidents in microblogs,” in The Semantic Web: ESWC 2013 Satellite Events, vol. 7955. Berlin, Germany: Springer-Verlag, pp. 22–33.
- [22] TextBlob. TextBlob: Simplified Text Processing. <http://textblob.readthedocs.io/en/dev/>

# CHARACTER AND IMAGE RECOGNITION FOR DATA CATALOGING IN ECOLOGICAL RESEARCH

Shannon Heh

Lynbrook High School  
San Jose, California, USA

## **ABSTRACT**

*Data collection is an essential, but manpower intensive procedure in ecological research. An algorithm was developed by the author which incorporated two important computer vision techniques to automate data cataloging for butterfly measurements. Optical Character Recognition is used for character recognition and Contour Detection is used for image-processing. Proper pre-processing is first done on the images to improve accuracy. Although there are limitations to Tesseract's detection of certain fonts, overall, it can successfully identify words of basic fonts. Contour detection is an advanced technique that can be utilized to measure an image. Shapes and mathematical calculations are crucial in determining the precise location of the points on which to draw the body and forewing lines of the butterfly. Overall, 92% accuracy were achieved by the program for the set of butterflies measured.*

## **KEYWORDS**

*Computer Vision, Image Recognition, Character Recognition, Ecology, Butterfly Cataloging*

## **1. INTRODUCTION**

Data collection is an important step of scientific research, especially in ecological and evolutionary studies. Scientists must gather a large amount of data to perform analyses of certain species and support their hypotheses. Much of the information is still contained in physical books that include images and descriptions of each butterfly. The only way to digitize this data is to manually measure the body sizes of organisms from these images and type the information into documents and spreadsheets. However, this method is time-consuming and inefficient. It requires a lot of manpower and is also prone to inaccurate measurements due to human errors. The goal of my project is to employ computer vision techniques to facilitate mass data collection and measurement. With computer vision, the process can be sped up significantly.

This summer, during my Earth Science internship at Stanford University, one of my tasks was to go through 50 books, measure the forewing and body length of thousands of butterflies, extract taxonomic information about each species, and record everything in an Excel spreadsheet. The data collection process not only requires knowledge of Lepidoptera classification, but also needs focus and patience. Two interns can only measure 300 butterflies in three hours. My internship

experience inspired me to search for efficient ways to automate the data collection process. While studying computer vision with Professor Susan Fox, I realized the application of computer vision would be the perfect solution to optimize the data collection process for ecological research.

Two main techniques I investigated for this project are:

1. Optical Character Recognition (OCR): to read the taxonomic information of butterflies
2. Contour Detection: to process the images of butterflies

**Optical Character Recognition** is the conversion of typed, printed, or handwritten text by a computer into machine-readable text. I integrated an already existing OCR Engine called Tesseract to conduct the character identification. The butterfly descriptions and measurements are then scanned and input into columns of a spreadsheet using the *openpyxl* library in Python.

Measurement can accurately be done with computer vision. **Contour Detection** was necessary to isolate the shape of the butterfly and locate the points on the butterfly from which to measure their body and wing lengths. Although I used butterflies as a proxy for measurement, any object can be measured with computer vision. The following program eliminates the need for humans to make such measurements.

## 2. BACKGROUND INFORMATION

### 2.1. What is Tesseract?

Tesseract is an open-source Optical Character Recognition (OCR) engine that can detect approximately 100+ languages [1]. In this project, pyTesseract [2], a Python wrapper for Google's Tesseract-OCR Engine, is implemented for character detection. Tesseract follows a step-by-step process for character detection. The first step is a connected component analysis to organize character outlines into "Blobs." Text lines are then found and broken into words based on the spacing between words. Character recognition is described as a "two-pass process": the first pass involves recognizing each word and passing it to an adaptive classifier as training data, while the second pass recognizes words again to ensure that the all words are well-detected [3].

### 2.2. How are the butterflies measured?

Two measurements are taken to quantify the butterfly's body size. The first is the forewing length (or basal-apical length) which extends approximately from the upper half of the butterfly body to the outer edge of the butterfly's forewing as shown below. The second is the butterfly's body length which starts right below the butterfly's head and ends at the tip of the body. (In my program, the body length will be measured from the *top* of the head to the end of the body.)

### 2.3. How are the butterflies typically displayed in books?

Figure 2 is a mock-up example showing how the butterflies are displayed in books. The data recorded about the butterflies include but are not limited to: the book in which the butterflies are displayed, the butterfly's family, genus, species, subspecies, authority, authority year, sex, page number, figure number, magnitude or scale, forewing length, and body length. The characteristics

of the butterfly are often displayed beside or below the butterfly's images. Other descriptions of the butterfly's habitat and behaviors are sometimes included in the books, but unnecessary for data collection about the butterflies' body sizes.

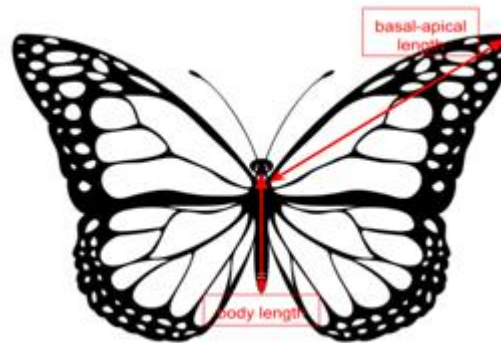


Figure 1. Measurements taken for a butterfly

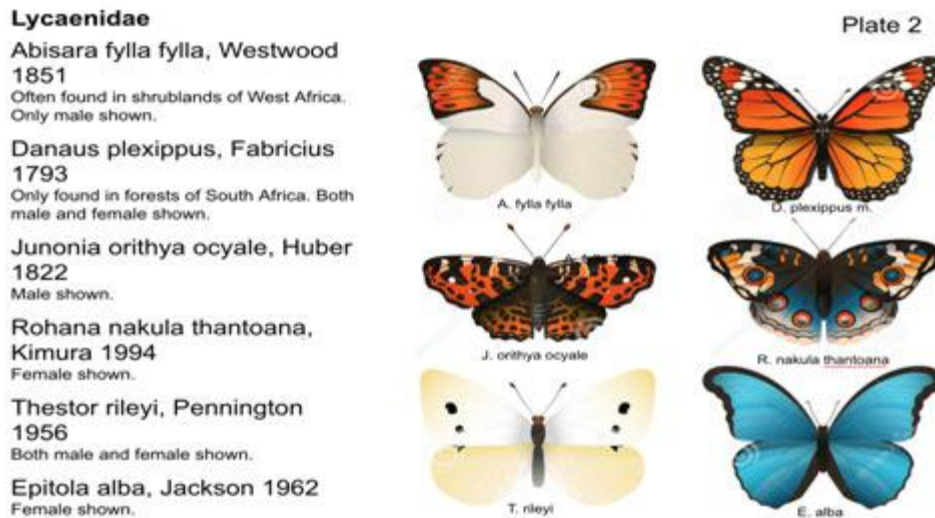


Figure 2. A mock-up example of a typical butterfly display in books

## 2.4. Related Work

In past studies of computer vision, butterflies were never detected and measured for ecological studies. However, fish have been studied and measured quite often to automate the inefficient manual sorting of fish. D.J. White, C. Svellingen, and N.J.C. Strachan [4] detected the principal axis of the fish by finding the two furthest points on the outline of a fish, an idea I implemented into my own project to determine the butterfly's forewing length.

Various interesting methods were proposed for text detection, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Neural Networks, and Tesseract for OpenCV. SVM, KNN, and Neural Networks all implement machine learning techniques to train the system to recognize letters and symbols. SVM [5] is a discriminative classifier defined by a hyperplane that maximizes

the margin of training data to find the best match between the characters with which the classifier has been trained and the character needing recognition. Yafang Xue [6] implemented SVM classifiers as one method of OCR, but the algorithm accuracy reached only 83%. P. Kumar, N. Sharma, and A. Rana [7] measured a 94.8% accuracy for the SVM Classifier, but only 80.96% accuracy for Neural Networks. KNN [8] is another classification algorithm which locates the “nearest neighbor,” or most similar match, in the training data to the character needing recognition. The data collector can gauge how many,  $k$ , neighbors should be detected for matching. While SVM and KNN are simple machine learning algorithms for character recognition, the process for training the classifiers is arduous; one needs to train the system for each character and font to ensure the highest accuracy. After reading a study from K.M. Sajjad [9] about Tesseract in automatic license plate recognition, I decided that using Tesseract in my project would be the fastest, most accurate OCR method, since the character recognition is already implemented into the engine.

The solution is divided into two phases: (1) character recognition and (2) butterfly measurement and detection. Phase one incorporates Optical Character Recognition (OCR) to recognize and print the characters in the image, while phase two manipulates various Python functions and libraries to detect contours, measure linear distances on butterflies, and enter data into a spreadsheet. The two-phase solution can be applied to scientific research for automated and efficient data collection.

### 3. CHARACTER RECOGNITION

#### 3.1. Pre-processing

Pre-processing is necessary to enhance the features of an image. Listed below are the pre-processing techniques that I implemented.

- 1) Gray-scale: The first step is to convert the image to a gray-scale image which only uses one channel and eliminates extraneous color information. Feature detection is improved and noise is reduced as a result.
- 2) Binarization: Binarization converts the image to black and white using a threshold value. Binary thresholding sets values above the threshold to the max value (white), and values less than or equal to the threshold to zero (black).
- 3) Resize: The next step is to resize the dimensions of the image to two times the original. By testing the code, I found that scaling the text larger makes the character recognition more accurate. The font size is assumed to be around 10-12 points.
- 4) Erosion: The following pre-processing step is erosion, which adds a layer of pixels to the image and thickens the text.

Sharpening: The final pre-processing step is sharpening. Sharpening is used to enhance the edges and features of each character. For example, I found that with one text, the program would confuse the letter ‘l’ and the symbol ‘j’. After image sharpening, the program was able to make the distinction.



Figure 3. Result of each step of pre-processing

### 3.2. Optical Character Recognition

In the second half of the character recognition process, the program utilizes Tesseract, an open-source Optical Character Recognition (OCR) engine. Developed by HP in 1985, Tesseract is currently one of the most accurate OCR software tools.

### 3.3. Test the Program

I tested the program with three different fonts: sans-Serif (Calibri), Serif (Times New Roman), and Lucida Calligraphy to analyze the different ways the type of font can affect the accuracy. I also included narrow (Arial Narrow), italicized, bold, all capitalized, colored, and highlighted text, the full alphabet, and a text comparison for further analysis.

The standard text I use for the tests is:

Species: Danaus  
Genus: Plexippus  
Authority: Shannon 2017

Shown in the table below are the results for each font type and characteristic added to the text. Note that kerning and line spacing are detected by Tesseract. All text was written in size 12 font. Serif font was written in Times New Roman and Sans-Serif was written in Calibri.

In general, simple fonts (Times New Roman, Calibri, and Arial) are more accurately recognized by the program. Text written in all capitals produced more incorrect results compared to text written regularly. Sans-Serif fonts, written regularly, were all recognized by Tesseract. When the Sans-Serif text is written in all capital letters, the software produced some errors (e.g. in the word 'Plexippus').” Serif fonts, written both ways, are accurately detected. There is again only one small error in the word “Plexippus,” when written in all capitals. Recognition of fonts with ornamentation, such as Lucida Calligraphy, produced the most inaccurate results, with no words being detected correctly. Narrow spacing between letters (Arial Narrow) did not hinder the program’s detection; this text was actually detected with the best accuracy.

<b>Sans-Serif – regular</b> Species: Danaus Genus: Plexippus Authority: Shannon 2017	<b>Sans-Serif -- all capitalized</b> SPECIES: DANAUS GENUS: PLB`IPPUS AUTHORITY: SHANNON 2017
<b>Serif – regular</b> Species: Danaus Genus: Plexippus Authority: Shannon 2017	<b>Serif -- all capitalized</b> SPECIES: DANAUS GENUS: PLEXEPPUS AUTHORITY: SHANNON 2017
<b>Lucida calligraphy – regular</b> s_pecies: Damn: germs: ?{éxg'gpm Autfiority: Sfianmm 2017	<b>Lucida calligraphy -- all capitalized</b> S?ECITS.' DANAHS GEM: Ems AHWORIW: SWN 2017
<b>Arial narrow – regular</b> Species: Danaus Genus: Plexippus Authority: Shannon 2017	<b>Arial narrow -- all capitalized</b> SPECIES: DANAUS GENUS: PLEXIPPUS AUTHORITY: SHANNON 2017
<b>Italicized -- Sans-Serif, regular</b> Name: Danaus Plexippus	<b>Bold -- Sans-Serif, regular</b> Name: Danaus Plexlppus
<b>Full Alphabet -- Serif</b> ABCDEFGHIJKLMNOPQRSTUVWXYZ abcdefghijklmnopqrstuvwxyz	<b>Full Alphabet -- Sans-Serif</b> ABCDEFGHIJKLMNOPQRSTUVWXYZ ahodefgghijklmnopqrstuvwxyz
<b>Red Text -- Sans-Serif</b> Species: Danaus	<b>Yellow Highlight -- Sans-Serif</b> Genus: Plexippus

Figure 4. Final results of OCR in the program

I input the full alphabet to see whether the program could detect all letters with the two most basic Serif and Sans-Serif fonts. Both fonts' recognition produces the same results, with both c's replaced with o's. I then tested the program with italicized, bolded, colored, and highlighted text. Italicized and bolded text had no effect on the program, beside the small error in "Plexippus" where 'l' was confused with 'i' for the bold text. This may be because the bolding over-thickened the characters (which were already thickened during the eroding during pre-processing). The color and background color of the text did not affect the program's character detection because color information was eliminated in the pre-processing.

## 4. BUTTERFLY RECOGNITION AND MEASUREMENT

### 4.1. Pre-processing

The butterfly image was converted to a gray-scale image to eliminate unnecessary color information. Again, color information makes the image contours more difficult to detect. The image is also sharpened to make contours clearer.



## 4.2. Contour Detection

All the contours on the butterfly image are detected. The outer shape of the butterfly is needed, so only the second largest contour (by area) is chosen. This step is shown in Figure 5.

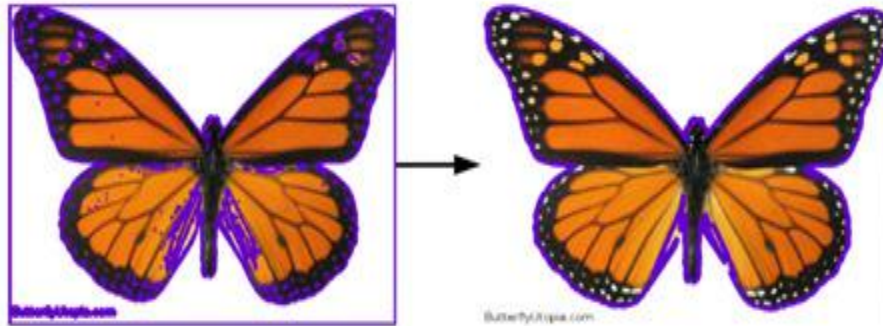


Figure 5. Before and after the largest contour was found

## 4.3. Measuring the Body Length

The following step is to calculate the body length. My initial approach was to locate the midline of the bounding box of the butterfly and find the highest and lowest points within 10 pixels of the midline, which correspond to the head and bottom end of the body. However, this method is ineffective for non-symmetrical butterflies, where the midline is skewed.

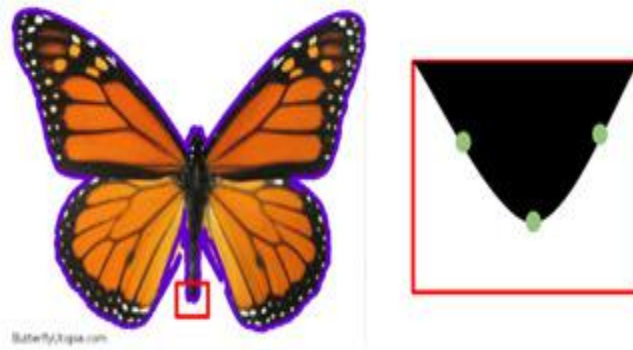


Figure 6. Close-up of the bottom end of the body

My current technique is to locate the bottom end of the butterfly body and calculate its distance from the head. I limited the range of contours to only look at the those within 50 pixels, left and right, of the midline to account for the potential skewness caused by non-symmetrical cases. Next, I chose three points that are 0, 2, and 4 pixels apart from the first. I then tested whether the middle point is the local minima as displayed in Figure 7. A line is directly drawn from the bottom end to the point where the line and head intersect, and the body length is determined.



Figure 7. Body length of the butterfly (in green)

#### 4.4. Measuring the forewing length

##### 4.4.1 Drawing an Ellipse

To find the butterfly's forewing length, I first drew an ellipse, which is similar to the butterfly body shape, around the butterfly's body. The major axis is half of the body length (Figure 8) and the minor axis is around 10 pixels, an approximation of the maximum body width of the butterfly. The minor axis is kept constant because there is usually little variation in the body width of butterflies among different species.



Figure 8. Ellipse around the butterfly's body (in red)

##### 4.4.2. Locating Point A

The next task is to find the two points on which to draw the line for the forewing length. Figure 9 shows where the points are usually found. I will call one point A, and the other point B.

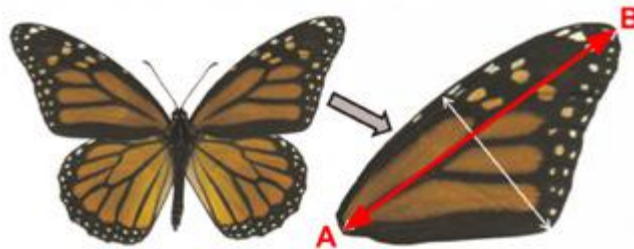
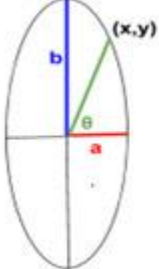


Figure 9. How the butterfly's forewing length is measured

I tested two different methods for detecting Point A. I first tried detecting the top of the butterfly head, then shifted this point 10 pixels down and five pixels to the right, and used that point as Point A. However, because of variations in the butterfly's body size, this method proved unreliable. After drawing the ellipse, I discovered that I could use the equation of the ellipse to find Point A(x, y) on the body.



$$x = \frac{ab}{\sqrt{a^2 \tan^2 \theta + b^2}} \quad y = x \cdot \tan \theta \quad (1)$$

After testing multiple angles for  $\theta$ , I observed that  $\theta$  should be around  $\pi/20$ .

#### 4.4.3 Locating Point B

The last step for the wing length measurement is finding Point B. I predicted that Point B would be the furthest point on the butterfly contour from Point A. Point B, in this case, must be on the right wing, so I made sure that the point was above and to the right of Point A. Finally, the distance formula was used to calculate the distance between Point A and B.



Figure 10. Forewing length of the butterfly (in green)

#### 4.5. Converting from Pixels to Millimeters

The final task for butterfly measurements is converting the body and wing length units from pixels to millimeters. I measured both lengths on a printed image of a butterfly, and used the pixel to millimeter ratio as the scale.

#### 4.6. Separating the butterfly image from the text

The contours around the letters can be a distraction to the program's line and contour detection on the butterfly. To avoid this issue, I separated the butterfly and text into two sub-images. Anything below the lowest point on the butterfly contour would be considered as text (assuming the text is below the butterfly). The character recognition was done on the text image, while butterfly measurements were performed on the butterfly images.

#### 4.7. Sorting the information into an Excel Spreadsheet

I incorporated the *openpyxl* library in Python to manipulate Excel spreadsheets for data collection. Each time a set of images is passed through the program, a new workbook is generated. The first row of the spreadsheet are the designated headers, as shown in Figure 12. Labels in original image are used for each line of text to identify the genus, species, sex, etc. (in reality, there are no labels for the butterfly's description). Each line of text is then separated into a list of strings. For each butterfly image, a row is appended to the worksheet, and the corresponding information is filled in the cells by matching the text labels to the column headers. An example is shown in Figure 12.

The text labels are detected through fuzzy matching, with the *SequenceMatcher* function in the *difflib* library. In case Tesseract incorrectly recognizes some character(s) in the text label, I set the standard so that as long as 75% of the characters in the label are correct, the program can match the label with the header.

**Family** → YES (100% match)   **Famly** → YES (83% match)   **Femllv** → NO (50% match)

#### 4.8. Test the Program

The Python program was tested on 12 images of butterflies taken from the book *Learning About Butterflies* by Carolyn Klass and Robert Dirig [10] and various websites. The program goes through all given directories and subdirectories to select the butterfly images.

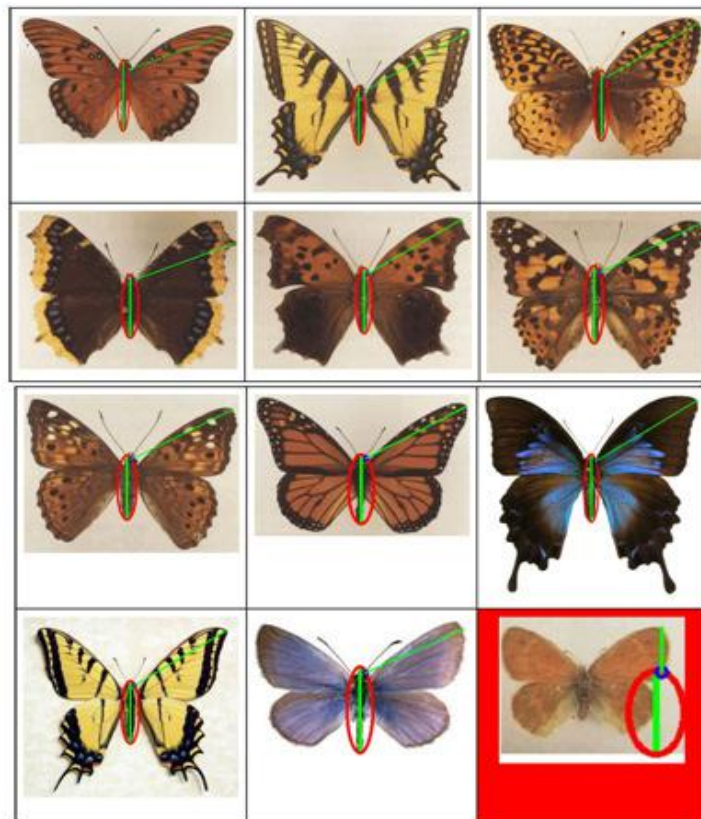
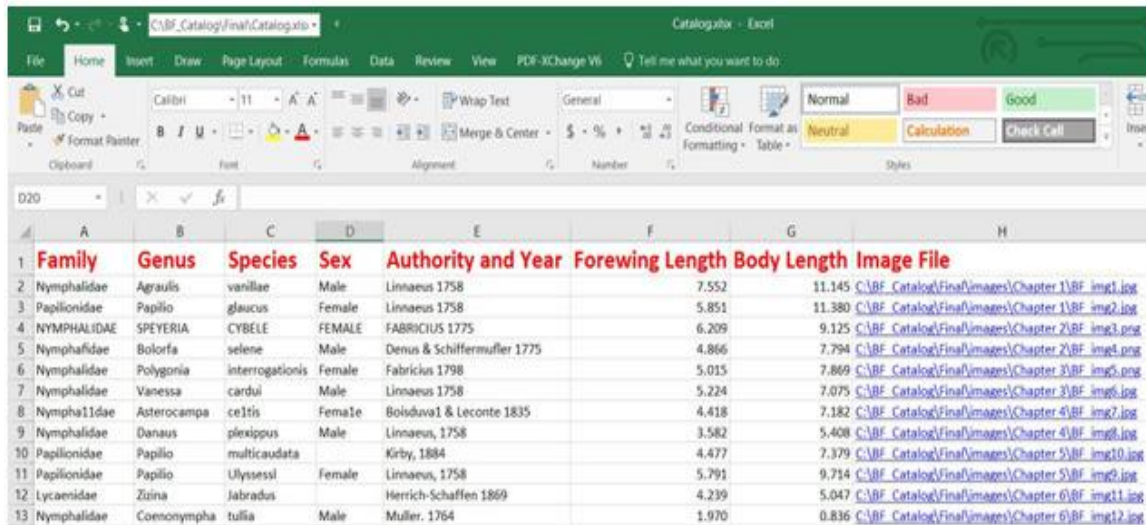


Figure 11. Final result for the 12 butterfly images

Results show that 11 out of 12 (92%) butterflies were successful in being measured and cataloged. The reason that the one butterfly (red box in Figure 14) was not identified correctly is that the butterfly's wing color is similar to the background color. These factors affected the program's ability to distinguish the contour of the butterfly. I also noticed that the rotation of the butterfly image disrupts the program's ability to detect the contours. Because of differences in font for the text, there were minor errors in the character recognition. Overall, this program is proven to be highly accurate and efficient for measurements and data cataloging.



	A	B	C	D	E	F	G	H
	Family	Genus	Species	Sex	Authority and Year	Forewing Length	Body Length	Image File
1	Nymphalidae	Agraulis	vanillae	Male	Linnaeus 1758	7.552	11.145	C:\BF_Catalog\Final\Images\Chapter 1\BF_img1.jpg
2	Papilionidae	Papilio	glaucus	Female	Linnaeus 1758	5.851	11.380	C:\BF_Catalog\Final\Images\Chapter 1\BF_img2.jpg
3	Nymphalidae	SPEYERIA	CYBELE	FEMALE	FABRICIUS 1775	6.209	9.125	C:\BF_Catalog\Final\Images\Chapter 2\BF_img3.png
4	Nymphalidae	Boloria	selenae	Male	Dennis & Schiffmuller 1775	4.866	7.794	C:\BF_Catalog\Final\Images\Chapter 2\BF_img4.png
5	Nymphalidae	Polygona	interrogationis	Female	Fabricius 1798	5.015	7.869	C:\BF_Catalog\Final\Images\Chapter 3\BF_img5.png
6	Nymphalidae	Vanessa	cardui	Male	Linnaeus 1758	5.224	7.075	C:\BF_Catalog\Final\Images\Chapter 3\BF_img6.jpg
7	Nymphalidae	Asterocampa	celtis	Female	Boldruga & Leconte 1835	4.418	7.182	C:\BF_Catalog\Final\Images\Chapter 4\BF_img7.jpg
8	Nymphalidae	Danaus	plexippus	Male	Linnaeus, 1758	3.582	5.408	C:\BF_Catalog\Final\Images\Chapter 4\BF_img8.jpg
9	Papilionidae	Papilio	multicaudata		Kirby, 1884	4.477	7.379	C:\BF_Catalog\Final\Images\Chapter 5\BF_img9.jpg
10	Papilionidae	Papilio	Ulysses	Female	Linnaeus, 1758	5.791	9.714	C:\BF_Catalog\Final\Images\Chapter 5\BF_img10.jpg
11	Lycenidae	Zizina	Jabradus		Herrich-Schaffner 1869	4.239	5.047	C:\BF_Catalog\Final\Images\Chapter 6\BF_img11.jpg
12	Nymphalidae	Comonympha	tullia	Male	Muller, 1764	1.970	0.836	C:\BF_Catalog\Final\Images\Chapter 6\BF_img12.jpg

Figure 12. Cataloged data in spreadsheet for 12 butterflies

## 5. CONCLUSIONS

Data collection is an essential, but a time-consuming and manpower intensive step in scientific research. In this project, two important computer vision methods were implemented to automate data cataloging for ecological research. Optical Character Recognition is an effective approach to scan in written or printed information, which will then be sorted into a spreadsheet. Proper pre-processing is first done on the images before Tesseract can be integrated for character recognition. Although there are limitations to Tesseract's detection of certain fonts, overall, Tesseract can successfully identify words of basic fonts. Contour detection is an advanced technique that can be utilized to measure an image. Shapes and mathematical calculations are crucial in determining the precise location of the points on which to draw the body and forewing lines of the butterfly. Finally, the butterfly information can be input into a spreadsheet. While this program is currently limited to butterfly measurements, similar techniques can be applied to the measurement of more organisms.

With the help of computer vision, scientists no longer need to invest significant amounts of their time on data cataloging and measurement. The outcome of this project allows researchers to automate the collection process and focus more time on their research and analyses.



## ACKNOWLEDGEMENTS

I would like to thank Professor Susan Fox at Macalester College for mentoring me in computer vision and guiding me through this research project. Thank you to the staff at Pioneer for all your help. Thank you to my Earth Science internship mentor Dr. Noel Heim at Stanford University for your encouragement in pursuing this project.

## REFERENCES

- [1] Tesseract OCR (Optical Character Recognition) Google. Retrieved September 04, 2017, from <https://opensource.google.com/projects/tesseract>
- [2] pytesseract 0.1.7 Python Package Index. Retrieved September 04, 2017, from <https://pypi.python.org/pypi/pytesseract>
- [3] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on (Vol. 2, pp. 629-633). IEEE.
- [4] White, D. J., Svellingen, C., & Strachan, N. J. C. (2006). Automated Measurement of species and length of fish by computer vision. Fisheries Research, 80(2), 3rd ser., pg. 203-210.
- [5] Introduction to Support Vector Machines (SVM). Retrieved September 2017 [http://docs.opencv.org/2.4/doc/tutorials/ml/introduction\\_to\\_svm/introduction\\_to\\_svm.html](http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html)
- [6] Xue, Y. (2014). Optical Character Recognition. Department of Biomedical Engineering, University of Michigan. <https://pdfs.semanticscholar.org/5b6b/e3357dbdbac38e92515a7b7aebb7e622f635.pdf>
- [7] Kumar, P., Sharma, N., & Rana A. (2012). Handwritten Character Recognition using Different Kernel based SVM Classifier and MLP Neural Network (COMPARISON) International Journal of Computer Applications, Volume 53, No. 11, September 2012.
- [8] Understanding k-Nearest Neighbour. Retrieved September 04, 2017, from [http://docs.opencv.org/3.0beta/doc/py\\_tutorials/py\\_ml/py\\_knn/py\\_knn\\_understanding/py\\_knn\\_understanding.html#knn-understanding](http://docs.opencv.org/3.0beta/doc/py_tutorials/py_ml/py_knn/py_knn_understanding/py_knn_understanding.html#knn-understanding)
- [9] Sajjad, K. M. (2012). Automatic license plate recognition using python and opencv. Department of Computer Science and Engineering MES College of Engineering. Retrieved Sept., 2017 [http://sajjad.in/content/ALPR\\_paper.pdf](http://sajjad.in/content/ALPR_paper.pdf)
- [10] Dirig, R. & Klass, C. (1992, March) Learning About Butterflies. Cornell University, NY: Cornell Cooperative Extension.
- [11] Sweigart A. (2015, April 14) Automate the Boring Stuff with Python: Practical Programming for Total Beginners. San Francisco, CA: No Starch Press, Inc.

## AUTHOR

Shannon Heh is a high school student (11th grade) at Lynbrook High School, San Jose, California, USA. She is the president of Girls Who Code organization of San Jose region in 2018 to 2019.



# PROBABILITY BASED CLUSTER EXPANSION OVERSAMPLING TECHNIQUE FOR IMBALANCED DATA

Shaukat Ali Shahee and Usha Ananthakumar

Shailesh J. Mehta School of Management,  
Indian Institute of Technology Bombay, Mumbai, India

## ABSTRACT

*In many applications of data mining, class imbalance is noticed when examples in one class are overrepresented. Traditional classifiers result in poor accuracy of the minority class due to the class imbalance. Further, the presence of within class imbalance where classes are composed of multiple sub-concepts with different number of examples also affect the performance of classifier. In this paper, we propose an oversampling technique that handles between class and within class imbalance simultaneously and also takes into consideration the generalization ability in data space. The proposed method is based on two steps- performing Model Based Clustering with respect to classes to identify the sub-concepts; and then computing the separating hyperplane based on equal posterior probability between the classes. The proposed method is tested on 10 publicly available data sets and the result shows that the proposed method is statistically superior to other existing oversampling methods.*

## KEYWORDS

*Supervised learning, Class Imbalance, Oversampling, Posterior Distribution*

## 1. INTRODUCTION

Class imbalance is one of the most challenging problems in Data Mining [1]. It refers to data sets where one class is under represented compared to another class also referred to as between class imbalance. This phenomenon is commonly seen in many real life applications like fault detection, fraud detection, anomaly detection, medical diagnosis [2][3][4]. Traditional classifiers applied to such imbalanced data fail to classify minority class examples correctly due to its bias towards majority class [5][6][7]. Owing to the large number of potential applications of class imbalance, various methods have been proposed in the literature to address this problem. These methods can be classified into four categories: Sampling based methods, Cost-sensitive learning, kernel-based learning and active learning. Though various approaches exist in literature to handle class imbalance problem, sampling based methods have shown great potential as they attempt to improve data distribution rather than the classifier [8][9][10][11]. Liu et al. [12] have given a number of reasons on why sampling methods are preferred compared to other methods.

Sampling based method is a pre-processing technique that diminishes the class imbalance effect either by increasing the minority class examples or by decreasing the majority class examples

[13][14]. In this study, we focus on oversampling as undersampling of majority class is not recommended when the dataset has absolute rarity of minority class [15]. In case of oversampling, the number of minority class examples is increased either by random replication of examples or by generating new synthetic examples to minimize the overfitting problem. With regard to synthetic sampling, the synthetic minority over-sampling technique (SMOTE) [8] generates synthetic minority class examples. The method first randomly selects a minority class example and then chooses its  $k$  nearest neighbours belonging to the minority class. Synthetic examples are generated between the example under consideration and the selected nearest neighbour example along the line joining them. However, while selecting the nearest neighbours of the minority class, it does not consider the majority class examples and gives equal weight to all nearest neighbours. Oversampling the examples along the line joining the considered example and the selected nearest neighbour leads to the problem of overlapping between the classes [16]. Adaptive synthetic sampling approach for imbalanced learning (ADASYN) [17] adaptively generates synthetic minority class examples based on their weighted distribution of minority class examples according to the level of difficulty in learning. This method generates more synthetic examples corresponding to hard to learn examples and less synthetic instances corresponding to easier to learn examples. Thus the method reduces the bias due to class imbalance and adaptively shifts the classification decision boundary towards hard to learn examples. The crux of the method is to identify hard to learn minority class examples and ADASYN sometimes fails to find the minority class examples that are closer to the decision boundary [9]. Majority weighted minority oversampling technique (MWMOTE) [9] is effective in selecting hard to learn minority class examples but in this method, small concepts present in minority class examples that are located far from majority class examples are not identified. For handling this problem which is also referred to as within class imbalance in literature, various cluster based methods have been proposed in literature [18][10][19]. Cluster Based Oversampling (CBO) [19] is an oversampling technique that can handle between-class imbalance and within-class imbalance simultaneously. However, this method uses random oversampling to oversample the sub-clusters and thus could result in the problem of overfitting.

Further, though class imbalance has been studied well in literature, the simultaneous presence of between class imbalance and within class imbalance has not been addressed enough. In this paper, we propose a method that can reasonably handle between class imbalance and within class imbalance simultaneously. It is an oversampling approach and also considers the generalization ability of the classifier. We have validated our proposed method on publicly available data sets using neural network and compared with existing oversampling techniques that rely on spatial location of minority class examples in the Euclidean feature space.

The remainder of the paper is divided into three sections. Section 2 discusses the proposed method and its various components. Analysis on various real life data sets is presented in Section 3. Finally, Section 4 concludes the paper with future work.

## 2. THE PROPOSED METHOD

The main objective of the proposed method is in enabling the classifier to give equal importance to all the sub-clusters of the minority class that would have been otherwise lacking due to skewed distribution of the classes. The other objective is to increase the generalization ability of the classifier on the test dataset. Generally, the classifier tries to minimize the total error and when the class distributions are not balanced, minimization of total error gets dominated by



minimization of error due to majority class. Neural network is one such classifier that minimizes the total error. The first objective is achieved by removal of between class and within class imbalance as it helps the classifier in giving equal importance to all the sub clusters. The second objective is realized by enlarging the data space of the sub-clusters as it increases the generalization ability of the classifier on test set.

In the proposed method, the first step is to normalize the input dataset between [0, 1] and then to remove the noisy examples from the dataset. A noisy example is identified based on K-Nearest Neighbour (KNN) of the considered example. In our method, we consider an example to be noisy if it is surrounded by 5 examples of the other class as also being considered in other studies including [9]. Removal of noisy examples helps in reducing the oversampling of noisy examples. After the removal of noisy examples, the concept present in data is detected using model based clustering. The boundary of the sub-clusters is computed based on the equal posterior probability of the classes. Subsequently, the number of examples to be oversampled is determined. Following subsections elaborate the proposed method in detail.

### 2.1. Locating each sub-concept

Model based clustering is used with respect to the classes to identify the sub-clusters (or sub-concepts) present in the dataset [20]. Model based clustering assumes that data are generated by a mixture of probability distributions in which each component corresponds to a different cluster. We have used MCLUST [21] for implementing the model based clustering. MCLUST is an R package that implements the combination of hierarchical agglomerative clustering, Expectation Maximization (EM) and the Bayesian Information criterion (BIC) for comprehensive cluster analysis.

### 2.2. Locating the separating hyperplane between the sub-clusters

Model based clustering assumes that data comes from mixture of underlying probability distributions in which each component represents a different group or cluster. In general it considers mixture of multivariate normal distributions. In computing the separating hyperplane between sub-clusters of two classes, the majority class sub-cluster is identified on the basis of the nearest neighbour examples of the minority class sub-cluster. A separating hyperplane is then computed between these two sub-clusters where the posterior probability between these two classes are considered equal. We have

$$p(y = 1|x) = p(y = 0|x) \dots\dots\dots (1)$$

which is same as

$$p(x|y = 1)p(y = 1) = p(x|y = 0)p(y = 0) \dots\dots\dots (2)$$

As oversampling handles between class and within class imbalance thus making prior probability equal, equation (2) reduces to

$$p(x|y = 1) = p(x|y = 0) \dots\dots\dots (3)$$

Since we assume that distribution is multivariate normal, the final equation of separating hyperplane is

$$(\mu_1 - \mu_2)^T \Sigma^{-1} x = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 + \mu_2) / 2 \quad (4)$$

where  $x \in R^n$ ,  $n$  is the number of features of the dataset;  $\mu_1$  is the mean of the minority class sub-cluster and  $\mu_2$  and  $\Sigma$  are respectively the mean and covariance of the majority class sub-cluster.

After computing the separating hyperplane between the sub-clusters, we expand the size of sub-clusters till the boundary of the region given by the hyperplane while maintaining the same structure as shown in Figure 1.

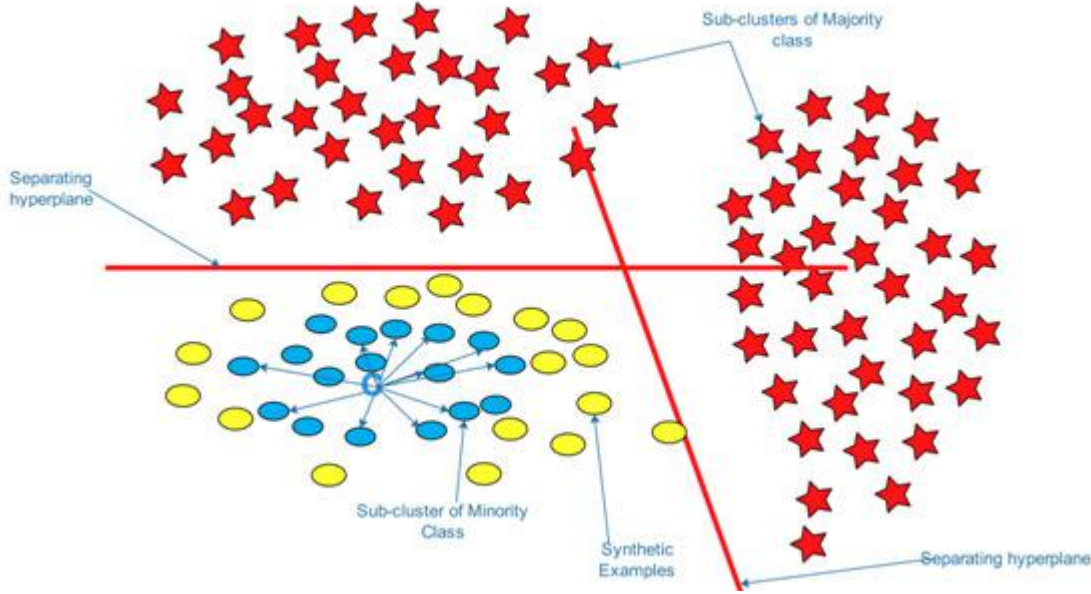


Figure 1. Illustration of synthetic examples being generated in the enclosed region of separating hyperplane

Maintaining the same structure can be done by multiplying the data points by a scalar quantity. The angle between the two vectors are defined as

$$\cos \theta = \langle v_1, v_2 \rangle / (|v_1| |v_2|) \quad (5)$$

If we multiply the vectors by a scalar  $\alpha$

$$\cos \theta = \langle \alpha v_1, \alpha v_2 \rangle / (|\alpha v_1| |\alpha v_2|) \quad (6)$$

As  $\cos \theta$  does not change, multiplying the vectors by a scalar does not change the structure of the data. The scalar  $\alpha$  is the minimum of the perpendicular distances from the centroid of the sub-cluster to the neighbouring separating hyperplanes. After computing  $\alpha$ , the synthetic example is generated by randomly selecting the minority class example  $u$  and extrapolating that example by using the following equation

$$\text{Synthetic Example} = C + (u - C)\alpha \quad (7)$$

In a situation where all the minority class examples of the sub-clusters lie outside the enclosing hyperplane of the sub-cluster as shown in Figure 2, synthetic examples have been generated inside the enclosing region of hyperplane using the following steps.

1. Let  $C$  be the centroid of the sub-cluster and vector  $u$  lie outside the region.
2. Substituting  $x = C + t(u - C)$  in equation (4), we get the  $t$  value.
3. Generate the uniform number between  $[0-t]$
4. *Synthetic Example* =  $C + t(u - C)$  ..... (8)

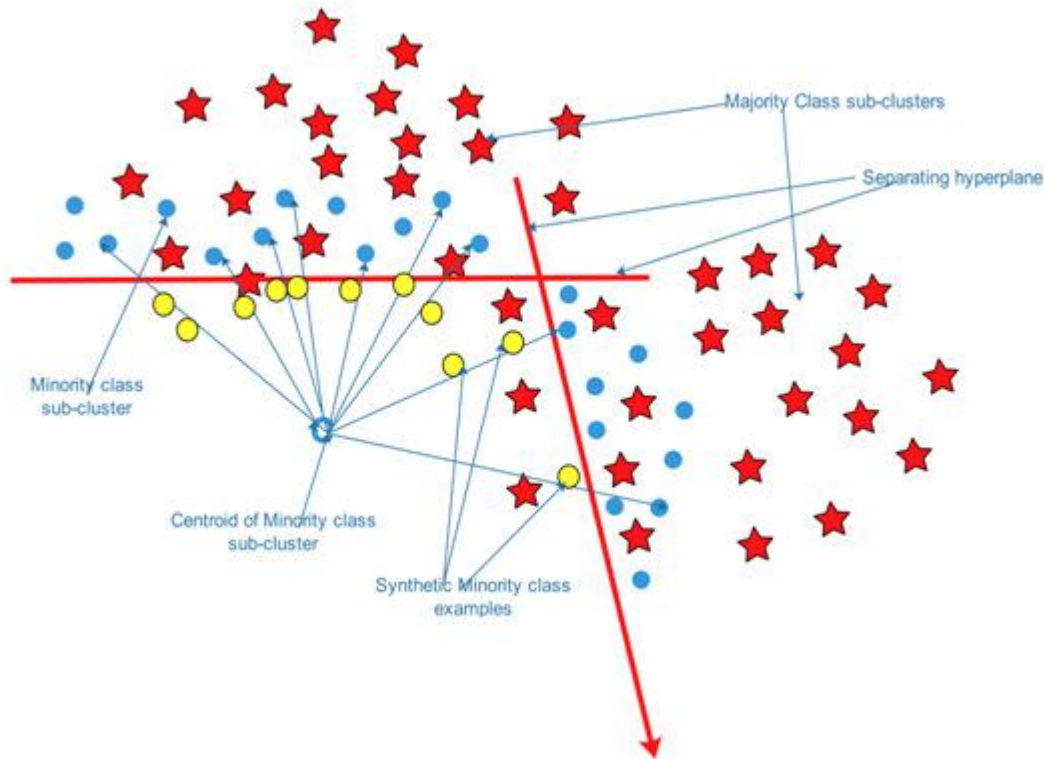


Figure 2. Illustration of synthetic examples being generated when all examples of the sub-cluster lie outside the enclosing hyperplane

### 2.3. Number of examples to be oversampled

After computing the enclosed region of the sub-clusters given by hyperplane, the method then computes the number of minority class examples to be oversampled. For this, we compute

$$T = T_{smaj}/q \text{ ..... (9)}$$

where  $T_{smaj}$  is the total number of majority class examples and  $q$  is the total number of sub-clusters of minority class. The number of examples to be oversampled in the  $i^{th}$  sub-cluster of minority class is given by

$$t_i = T - a_i \text{ ..... (10)}$$

where  $a_i$  is the number of examples already present in the  $i^{th}$  sub-cluster.

## 2.4. Algorithm

Input: Training dataset:

$$S = \{X_i, y_i\}, i = [1 - m]; X_i \in R^n \text{ and } y_i \in \{0,1\}$$

with Tsmaj = Total no of majority class example and Tsmi = Total no of minority class examples.

Output: Oversampled Dataset

1. Remove the noisy examples from the dataset
2. Applying model based clustering with respect to the classes gives
3. A set of  $q$  minority class sub-clusters  $\{smin_1 \dots \dots \dots smin_q\}$
4. A set of  $r$  majority class sub-clusters  $\{smaj_1 \dots \dots \dots smaj_r\}$
5. **For** each minority class sub-cluster  $\{smin_1 \dots \dots \dots smin_q\}$
6.     **For** each majority class sub-cluster  $\{smaj_1 \dots \dots \dots smaj_r\}$
7.         Compute the separating hyperplane between the sub-cluster as explained in section 2.2.
8.     **EndFor**
9. **EndFor**
10. # Oversampling the minority class sub-clusters
11.  $T = Tsmaj/q$
12. **For** each of the minority class subclusters  $\{smin_1 \dots \dots smin_q\}$
13.  $a_i = \text{size}(smin_i)$
14.  $t_i = T - a_i$
15. **If**  $smin_i$  lies completely outside the enclosed region
16.     **then** Generate synthetic examples using equation (8) of section 2.2
17. **Else**
18.     **While**  $(t_i > 0)$
19.          $S = \text{sample}(smin_i, t_i)$  # Select  $t_i$  examples from  $smin_i$
20.         Let  $s_i$  be the number of examples lying inside the enclosed region. Generate synthetic examples using equation (7) as explained in section 2.2
21.          $t_i = t_i - |s_i|$
22.     **EndWhile**
23. **EndElse**
24. **EndIf**
25. **EndFor**

## 3. COMPARATIVE ANALYSIS

In this section, we evaluate the performance of our proposed method and compare its performance with SMOTE [10], ADASYN [12], MWMOTE [13] and CBO [16]. The proposed method is evaluated on 10 publicly available data sets from KEEL [22] dataset repository. The data sets considered in this study are listed in Table 1.

### 3.1. Data sets

As this study is about binary classification problem, we have made modifications on yeast dataset as this is multiclass dataset, and the rest of the data sets were taken as it is. In case of yeast dataset, it has 10 classes {MIT, NUC, CYT, ME1, ME2, ME3, EXC, VAC, POX, ERL}. We chose ME3 as the minority class and the remaining classes were combined to form the majority class thus making it an imbalanced dataset. Table 1 represents the characteristics of various data sets used in this study.

Table 1. The Data sets.

Data sets	Total Examples	No. Minority Example	No Majority Exp	Attributes
glass1	214	76	138	9
pima	768	268	500	8
glass0	214	70	144	9
yeast1	1484	429	1055	8
vehicle2	846	218	628	18
ecoli1	336	77	259	7
yeast	1484	163	1321	8
yeast3	1484	163	1321	8
yeast-0-5-6-7-9 vs 4	528	51	477	8
yeast-0-2-5-7-9 vs 3-6-8	1004	99	905	8

### 3.2. Assessment metrics

Traditionally, performance of the classifier is based on accuracy and error measure that is defined as follows

$$Accuracy = \frac{(TP + TN)}{Total\ Examples}$$

$$Error\ rate = 1 - Accuracy$$

where TP is the number of positive examples classified correctly and TN is the number of negative class examples classified correctly. However, in case of imbalanced datasets, this accuracy measure overestimates the performance of the classifier as this measure could be high even when all or most of the minority class examples are misclassified. To deal with this problem, Haibo [11] proposed various alternative metrics based on the confusion matrix shown in Table 2.

Table 2. Confusion Matrix

		True Class	
Classifier Output		P	N
	P	TP	FP
	N	FN	TN

Some of the alternative measures are precision, recall, F-measure and G-mean defined as

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Measure = \frac{(1 + \beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision}$$

$$G - Mean = \sqrt{\frac{TP}{TP + FN} * \frac{TN}{TN + FP}}$$

Here  $\beta$  is a non-negative parameter that controls the influence between precision and recall. In this study we set  $\beta = 1$  implying that both are of equal importance. Another widely used accuracy measure is Receiving Operating Characteristic (ROC) curve that gives a graphical representation of classifier performance. The area under this curve is known as AUC measure. Since F-measure combines precision and recall, we provide F-measure, G-mean and AUC in this study.

### 3.3. Experimental setting

In this study, we use feed forward neural network with back propagation in order to evaluate the performance of the proposed method. This particular classifier is chosen in this study as it is one of the classifiers that minimizes the total error and the present algorithm by way of removing the between class and within class imbalance is expected to result in improved performance of the classifier. In the parameter setting, the number of input neurons being used is equal to the number of features and sigmoid function is being used as the activation function with learning rate 0.3. The number of hidden layers being used is one and the number of neurons it contains is equal to (number of features + classes)/2 [23].

Further, in SMOTE, the number of nearest neighbours used is set to 5. In case of ADASYN, we use  $k = 5$  and balance level = 1. In case of MWMOTE, we set the parameters as  $k1 = 5$ ,  $k2 = 3$ ,  $k3 = \text{lsmin}/2$ ,  $cp = 3$  and  $cf(th) = 5$ .

### 3.4. Results

The results of the 10 data sets are shown in Table 3 where in each measure, the maximum value has been highlighted. Stratified 5-fold cross validation technique was carried out where oversampling was carried out only in the training data containing four of the folds and the fifth fold was used as the test set. The model was trained on the oversampled training set and applied on the test data set in order to obtain an unbiased estimate of the model. This process was replicated five times and its average is presented in Table 3.

Results in Table 3 show that the proposed method performs better than the other methods in most of the data sets. It can be observed that the AUC value of the proposed method is better than other oversampling methods except *yeast1* dataset.

Table 3. F-Measure, G-Mean and AUC for 10 data sets.

Data	Method	F-Measure of Majority Class	F-Measure of Minority Class	G-Mean	AUC
glass1	SMOTE	0.745	0.619	0.690	0.721
	ADASYN	0.757	0.606	0.683	0.717
	MWMOTE	0.759	0.605	0.684	0.728
	CBO	0.760	<b>0.624</b>	<b>0.699</b>	0.736
	<b>Prop. Method</b>	<b>0.803</b>	0.616	0.694	<b>0.767</b>
Pima	SMOTE	0.750	0.631	0.707	0.759
	ADASYN	0.766	0.622	0.703	0.765
	MWMOTE	0.748	0.623	0.701	0.763
	CBO	0.766	0.608	0.691	0.748
	<b>Prop. Method</b>	<b>0.814</b>	<b>0.639</b>	<b>0.716</b>	<b>0.795</b>
glass0	SMOTE	0.817	0.681	0.762	0.832
	ADASYN	0.819	0.680	0.761	0.820
	MWMOTE	0.812	0.678	0.758	0.819
	CBO	0.797	0.661	0.743	0.795
	<b>Prop. Method</b>	<b>0.842</b>	<b>0.713</b>	<b>0.785</b>	<b>0.847</b>
yeast1	SMOTE	0.785	0.575	0.699	0.772
	ADASYN	0.754	0.582	0.705	0.772
	MWMOTE	0.772	<b>0.584</b>	<b>0.707</b>	<b>0.776</b>
	CBO	0.670	0.550	0.660	0.727
	<b>Prop. Method</b>	<b>0.808</b>	0.572	0.692	0.770
vehicle2	SMOTE	0.982	0.950	0.970	0.993
	ADASYN	0.980	0.942	0.963	0.989
	MWMOTE	0.981	0.946	0.967	0.993
	CBO	0.982	0.949	0.968	<b>0.994</b>
	<b>Prop. Method</b>	<b>0.985</b>	<b>0.957</b>	<b>0.974</b>	<b>0.994</b>

ecoli1	SMOTE	0.913	0.723	0.822	0.916
	ADASYN	0.900	0.719	0.839	0.903
	MWMOTE	0.914	0.736	0.839	0.916
	CBO	0.901	0.711	0.829	0.911
	<b>Prop. Method</b>	<b>0.938</b>	<b>0.786</b>	<b>0.854</b>	<b>0.937</b>
Yeast	SMOTE	0.965	0.737	0.870	0.943
	ADASYN	0.955	0.712	<b>0.898</b>	0.938
	MWMOTE	0.965	0.734	0.866	0.941
	CBO	0.950	0.689	0.892	0.935
	<b>Prop. Method</b>	<b>0.967</b>	<b>0.752</b>	0.881	<b>0.959</b>
yeast3	SMOTE	0.966	0.743	0.870	0.943
	ADASYN	0.952	0.695	0.886	0.930
	MWMOTE	0.966	0.742	0.866	0.938
	CBO	0.945	0.671	<b>0.887</b>	0.936
	<b>Prop. Method</b> left, ..	<b>0.968</b>	<b>0.759</b>	0.878	<b>0.951</b>
yeast-0-5-6-7-9 vs 4	SMOTE	0.939	0.484	0.694	0.804
	ADASYN	0.921	0.458	0.725	0.824
	MWMOTE	0.942	0.489	0.685	0.819
	CBO	0.923	0.475	<b>0.728</b>	0.830
	<b>Prop. Method</b>	<b>0.948</b>	<b>0.506</b>	0.684	<b>0.851</b>
yeast-0-2-5-7-9 vs 3-6-8	SMOTE	0.972	0.760	0.873	0.920
	ADASYN	0.949	0.649	0.868	0.913
	MWMOTE	0.975	0.768	0.858	0.929
	CBO	0.948	0.638	0.857	0.913
	<b>Prop. Metho</b>	<b>0.976</b>	<b>0.787</b>	<b>0.882</b>	<b>0.933</b>

To test the statistical difference between the proposed method and other existing oversampling methods, we have performed Wilcoxon signed-rank non-parametric test [24] on the metric measures F-measure of minority and majority class, G-mean and AUC. The null and alternative hypotheses are as follows:

H0: The median difference is zero.

H1: The median difference is positive.

The test statistic of the Wilcoxon Signed Rank Test is defined as  $W = \min(W+, W-)$  where  $W+$  is the sum of the positive ranks and  $W-$  is the sum of the negative ranks. As 10 data sets have been used to carry out the test, the  $W$  value at a significance of 0.05 should be less than or equal to 10 to reject the null hypothesis. The details of Wilcoxon Signed Rank Test for AUC measure between the proposed method and MWMOTE is given in Table 4. As we can see from this table that  $W+ = 52$ ,  $W- = 3$ ,  $W = \min(W+, W-) \Rightarrow W = 3$ , we reject the null hypothesis and conclude that the proposed method is better than MWMOTE in terms of AUC measure.



Table 4. Wilcoxon Signed Rank Test of AUC between the proposed method and MWMOTE.

Dataset	Proposed method	MWMOTE	Difference	Rank
glass1	0.767	0.728	0.039	10
Pima	0.795	0.763	0.032	8.5
glass0	0.847	0.819	0.028	7
yeast1	0.770	0.776	-0.006	3
vehicle2	0.994	0.993	0.001	1
ecoli1	0.937	0.916	0.021	6
Yeast	0.959	0.941	0.018	5
yeast3	0.951	0.938	0.013	4
yeast-0-5-6-7-9 vs 4	0.851	0.819	0.032	8.5
yeast-0-2-5-7-9 vs 3-6-8	0.933	0.929	0.004	2
W+ = 52, W- = 3, W = min(52, 3) = 3				

For space consideration, we present just the summary of Wilcoxon Signed Rank Test between the proposed method and other oversampling methods for various metric measures in Table 5. From this table, it can be seen that the proposed method is statistically significantly better than the other oversampling methods in terms of AUC and F-measure of both majority and minority class, although in case of G-mean, the proposed method does not seem to outperform the other oversampling methods. Though it is desirable that any algorithm performs well on all the measures, as stated in [25], AUC is a measure that is not sensitive to the distribution of the two classes thus making it suitable as a performance measure for the imbalanced problem.

Table 5. Summary of Wilcoxon signed rank test between our proposed method and other methods

Method	Proposed Method	Metric Measure
SMOTE	W+ = 55, W- = 0, W = 0 W+ = 52, W- = 3, W = 3 W+ = 36, W- = 19, W = 19 W+ = 53, W- = 2, W = 2	F-Measure of Majority class F-Measure of Minority class G-mean AUC
ADASYN	W+ = 55, W- = 0, W = 0 W+ = 53.5, W- = 1.5, W = 1.5 W+ = 32.5, W- = 22.5, W = 22.5 W+ = 54, W- = 1, W = 1	F-Measure of Majority class F-Measure of Minority class G-mean AUC
MWMOTE	W+ = 55, W- = 0, W = 0 W+ = 53, W- = 3, W = 3 W+ = 47.5, W- = 7.5, W = 7.5 W+ = 52, W- = 3, W = 3	F-Measure of Majority class F-Measure of Minority class G-mean AUC
CBO	W+ = 55, W- = 0, W = 0 W+ = 53.5, W- = 1.5, W = 1.5 W+ = 40, W- = 15, W = 15 W+ = 55, W- = 0, W = 0	F-Measure of Majority class F-Measure of Minority class G-mean AUC

## 4. CONCLUSION

In this paper, we have proposed a method that can handle between class imbalance and within class imbalance simultaneously. The proposed method applies model based clustering with respect to each of the classes to identify the sub-concepts present in the dataset. Then it computes the separating hyperplane that satisfies the equal posterior probability between the sub-concepts.

It then generates the synthetic examples while maintaining the structure of the original dataset in the enclosed region given by the hyperplane thus increasing the generalization accuracy of the classifier.

The proposed method has been evaluated on 10 publicly available data sets and the results clearly show that the proposed method increases the accuracy of the classifier. However, the limitation of the proposed method is that it gets influenced by the nearest majority class sub-clusters in the expansion of the minority sub-clusters which could be extended as future work. Another possible extension could be in modifying the computation of separating hyperplane by including majority class clusters that are located far from minority class clusters.

## REFERENCES

- [1] Q. Yang et al., "10 Challenging Problems in Data Mining Research," *Int. J. Inf. Technol. Decis. Mak.*, vol. 5, no. 4, pp. 597–604, 2006.
- [2] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern Recognit.*, vol. 46, no. 12, pp. 3460–3471, 2013.
- [3] S. García and F. Herrera, "Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy," *Evol. Comput.*, vol. 17, no. 3, pp. 275–306, 2009.
- [4] S. Vajda and G. A. Fink, "Strategies for training robust neural network based digit recognizers on unbalanced data sets," *Proc. - 12th Int. Conf. Front. Handwrit. Recognition, ICFHR 2010*, no. November 2010, pp. 148–153, 2010.
- [5] S. Maldonado and J. Lopez, "Imbalanced data classification using second-order cone programming support vector machines," *Pattern Recognit.*, vol. 47, no. 5, pp. 2070–2079, 2014.
- [6] D. J. Yu, J. Hu, Z. M. Tang, H. Bin Shen, J. Yang, and J. Y. Yang, "Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling," *Neurocomputing*, vol. 104, pp. 180–190, 2013.
- [7] C. Y. Yang, J. S. Yang, and J. J. Wang, "Margin calibration in SVM class-imbalanced learning," *Neurocomputing*, vol. 73, no. 1–3, pp. 397–411, 2009.
- [8] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, pp. 321–357, 2002.
- [9] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning," *Knowl. Data Eng. IEEE Trans.*, vol. 26, no. 2, pp. 405–425, 2014.
- [10] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5718–5727, 2009.
- [11] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: A New Over-Sampling Method in," pp. 878–887, 2005.
- [12] A. Liu, J. Ghosh, and C. E. Martin, "Generative Oversampling for Mining Imbalanced Datasets," *Int. Conf. data Min.*, pp. 66–72, 2007.

- [13] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004.
- [14] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Comput. Intell.*, vol. 20, no. 1, pp. 18–36, 2004.
- [15] G. M. Weiss, "Mining with Rarity: A Unifying Framework," *SIGKDD Explor.*, vol. 6, no. 1, pp. 7–19, 2004.
- [16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [17] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks*, 2008, no. 3, pp. 1322–1328.
- [18] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "DBSMOTE: Density-based synthetic minority over-sampling technique," *Appl. Intell.*, vol. 36, no. 3, pp. 664–684, 2012.
- [19] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 40–49, 2004.
- [20] C. Fraley and a E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Am. Stat. Assoc.*, vol. 97, no. 458, pp. 611–631, 2002.
- [21] C. Fraley and A. E. Raftery, "MCLUST: Software for model-based cluster analysis," *J. Classif.*, vol. 16, no. 2, pp. 297–306, 1999.
- [22] J. Alcalá-Fdez et al., "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Log. Soft Comput.*, vol. 17, no. 2–3, pp. 255–287, 2011.
- [23] H. Guo and H. L. Viktor, "Boosting with Data Generation: Improving the Classification of Hard to Learn Examples.," *Iea/Aie*, vol. 3029, pp. 1082–1091, 2004.
- [24] A. Richardson, "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach by Gregory W. Corder, Dale I. Foreman," *Int. Stat. Rev.*, vol. 78, no. 3, pp. 451–452, 2010.
- [25] I. Nekooeimehr and S. K. Lai-Yuen, "Adaptive semi-supervised weighted oversampling (A-SUWO) for imbalanced datasets," *Expert Syst. Appl.*, vol. 46, pp. 405–416, 2016.

## AUTHORS

**Shaukat Ali Shahee** received the Bachelor's degree in Mathematics honors from Patna University and Master's degree in computer application from the West Bengal University of Technology, India. He is currently pursuing PhD at Shailesh J. Mehta School of Management, Indian Institute of Technology Bombay. His research interests include data mining, machine learning and applied statistics.



**Prof. Usha Ananthakumar** received PhD degree in Statistics from Indian Institute of Technology Bombay, India. She is currently a Professor at Shailesh J. Mehta School of Management, Indian Institute of Technology Bombay. Her research interests include data mining, machine learning, applied Statistics and multivariate data analysis.



# VALIDATION METHOD OF FUZZY ASSOCIATION RULES BASED ON FUZZY FORMAL CONCEPT ANALYSIS AND STRUCTURAL EQUATION MODEL

Imen Mguiris<sup>1</sup>, Hamida Amdouni<sup>2</sup> and Mohamed Mohsen Gammoudi<sup>3</sup>

<sup>1</sup>Computer Science Department, FST-University of Tunis ElManar,  
Tunis, Tunisia

<sup>2</sup>ESEN, University of Manouba, Manouba, Tunisia

<sup>3</sup>ISSAM, University of Manouba, Manouba, Tunisia

## ABSTRACT

*In order to treat and analyze real datasets, fuzzy association rules have been proposed. Several algorithms have been introduced to extract these rules. However, these algorithms suffer from the problems of utility, redundancy and large number of extracted fuzzy association rules. The expert will then be confronted with this huge amount of fuzzy association rules. The task of validation becomes fastidious. In order to solve these problems, we propose a new validation method. Our method is based on three steps. (i) We extract a generic base of non redundant fuzzy association rules by applying EFAR-PN algorithm based on fuzzy formal concept analysis. (ii) we categorize extracted rules into groups and (iii) we evaluate the relevance of these rules using structural equation model.*

## KEYWORDS

*Fuzzy Association Rules Validation, Fuzzy Formal Concept Analysis, Structural equation model*

## 1. INTRODUCTION

The extraction of association rules is one of the most known techniques of data mining [1]. It aims at discovering correlations between the properties (attributes) characterizing the objects saved in the databases. Correlations discovered allow decision-makers to make better judgments. Indeed, association rules have been used in several fields, including medical research [2], analysis of geographic data and biological data [3] and electronic commerce [4].

The integration of fuzzy logic into the extraction of association rules made it possible to solve the problem of discretization and to process the quantitative databases without loss of information. A fuzzy association rule was the object of several studies since the work of [5]. However, the major drawback of fuzzy association rule extraction algorithms is the large number of rules generated. As a result, it becomes very difficult to interpret and exploit these rules when making decisions. The expert is obliged to validate them manually. Several researchers have proposed various

methods of assisting evaluation in order to make this task less time-consuming. However, these problems still persist.

In this context, we present a new method of validation of fuzzy association rules exploiting the structural equations Model (SEM). Our method contains three steps. The first step consists of applying EFAR-PN algorithm to extract generic bases of association rules. These bases contain a set of non-redundant association rules. This algorithm is based on Fuzzy Formal Concepts Analysis. The second step consists of categorizing the extracted rules into groups based on their items. This step provides a synthetic representation of the rules. The Final step allows evaluating the rule by using Structural Equation Model (SEM). We are using specifically one of the SEM techniques known as Partial Least Square (PLS).

The remainder of this article is organized as follows. In section 2, we introduce some basic notions necessary to better understand our work. In section 3, we present the different categories of fuzzy association rules algorithms. Section 4 surveys related work. In section 5, we detail the principle of our method by illustrating it with an example. Section 6 is devoted to evaluate our method by performing a series of experiments on three test bases used by the scientific community of the field. Finally, section 6 presents a conclusion and some future work.

## 2. BASIC NOTIONS

In this section, we present some basic concepts related to our work [6][7]

- Association rules: an association rule is written in the following form:

$$R: A \rightarrow B$$

Where  $A \cap B = \emptyset$ . A is called the premise of the rule and B its conclusion. Two measures are used when extracting association rules:

- Support: It is the measure of the frequency of simultaneous appearance of an itemset AB in the set of objects, denoted  $\text{Supp}(AB)$ . An itemset is said to be frequent if its support is greater than or equal to a minimal support (minsup).
- Confidence: It is the probability of having the itemset B, knowing that we already have the itemset A. According to [8], this measure is equal to  $\text{Conf}(R: A \rightarrow B) = \text{Supp}(AB) / \text{Supp}(A)$ . An association rule is said to be valid if, and only if, its confidence is greater than or equal to the threshold set by the user called minconf.
- Fuzzy Formal Context: It is a triplet  $K = (O, I, R)$  describing a set of objects O, a set of attributes I and a fuzzy binary relation  $R \subseteq O \times I$ . The value  $u_R(o, i)$  with  $o \in O$  and  $i \in I$ , is the association degree between o and i.
- Fuzzy Galois Connection and Closure Operator:  $K = (O, I, R)$  is a fuzzy formal context, for  $X \subseteq O$  and  $Y \subseteq I$ . Operators  $\Phi$  and  $\Psi$  are defined as follows:  
The fuzzy operator  $\Phi$  is applied to a set of objects  $X \subseteq O$  to determine a fuzzy set of items associated with all objects of X having the minimal degree  
 $\Phi: P(O) \rightarrow P(I)$

$$\Phi(X) = \{i^\alpha \mid \forall o \in X, \alpha = \min \mu_R(o, i)\} \quad (1)$$

The fuzzy operator  $\Psi$  is applied on a fuzzy set of items  $Y \subseteq I$  providing a set of objects satisfying the constraint imposed by the input set.

$\Psi: P(I) \rightarrow P(O)$

$$\psi = \{o \mid \forall i, i \in Y, u_Y(i) \leq u_R(o, i)\} \quad (2)$$

- Fuzzy Minimal Generator: Let  $c$  be a fuzzy itemset,  $I'$  is FFCI, if  $I' = \phi(c)$  and  $\nexists c_1 \subseteq c$  such as  $\phi(c_1) = I'$ , then  $c$  is a minimal fuzzy generator of  $I'$ . It's frequent if its support is greater than minsup.
- Fuzzy Closed Itemset (FCI): An itemset  $I'$  is an FCI iff  $I' = \phi(I')$ . It's frequent if its support is greater than minsup.
- Partial order relation between concepts  $<<$ : Let  $(A_1, B_1)$  and  $(A_2, B_2)$  two fuzzy formal concepts:  $(A_1, B_1) << (A_2, B_2) \Leftrightarrow A_2 \subseteq A_1$  and  $B_1 \subseteq B_2$ .
- Meet/Join: For each pair of concepts  $(A_1, B_1)$  and  $(A_2, B_2)$ , there exists a greatest lower bound (resp. a least upper bound) called Meet (resp. Join) denoted as  $(A_1, B_1) \wedge (A_2, B_2)$  (resp.  $(A_1, B_1) \vee (A_2, B_2)$ ) and defined by

$$(A_1, B_1) \wedge (A_2, B_2) = (\Psi(B_1 \cup B_2), (B_1 \cup B_2)) \quad (3)$$

$$(A_1, B_1) \vee (A_2, B_2) = ((A_1 \cup A_2), \phi(A_1 \cup A_2)) \quad (4)$$

- Iceberg Lattice: It is a partially ordered structure of a frequent fuzzy closed itemset and having only a join operator.
- Fuzzy Equivalence Class: It is a set of frequent fuzzy itemsets having the same support and the same closure. The largest element of the equivalence class is a frequent fuzzy closed itemset called  $c$  and smaller ones are their minimal generators.
- Frequent Fuzzy Minimal Generators Lattice: It is a partially ordered structure where the nodes are equivalence classes.
- Generic base of exact fuzzy association rules

Generic base of exact associative rules (GBEF) is a base composed of non-redundant generic rules having a confidence ratio equal to 1 [9].

Let  $FG_k$  be the set of fuzzy frequent closed itemsets and  $FG_l$  the set of minimal generators of the itemset  $I$ . The generic base of exact fuzzy association rules is defined as follows:

$$GBEF = \{R : g \rightarrow (I - g) \mid I \in (FC_k) \wedge g \in (FG_l), g \neq I\}. \quad (5)$$

- Generic base of approximate fuzzy association rules

The generic base of approximate associative (GBAF) rules is defined as follows:

$$GBAF = \{R : g \rightarrow (I_1 - g) \mid I, I_1 \in FC_k, g \in FG_l \wedge I \subset I_1 \wedge conf(R) \geq minconf\}. \quad (6)$$

- Generic base of transitive fuzzy association rules

The generic base of Transitive associative (RIF) rules is defined as follows:

$$RIF = \{ R : g \rightarrow (I_1 - g) \mid I, I_1 \in FC_k, g \in FG_I \wedge I \subset I_1 \wedge \nexists I_2 \text{ s.t. } I \subset I_2 \subset I_1 \wedge \text{conf}(R) \geq \text{minconf} \}. \quad (7)$$

The exact fuzzy association rule is a relationship between the frequent fuzzy closed itemset FFCI and their minimal generators. However, the approximate rule is a relation of an FFCI with another FFCI that covers it and the transitive rule is a link between two FFCIs, one of which covers the other immediately.

### 3. FUZZY ASSOCIATION RULES

Association rule is one of the most important unsupervised methods of data mining also called Market Basket Analysis. Several algorithms have been proposed in the literature to extract association rules. These algorithms deal only with binary contexts, whereas the real databases include not only binary data, but also quantitative data. In order to apply these algorithms, the quantitative databases must be converted into binary bases. This transformation causes several problems, namely the loss of information. This loss causes the non-coverage of the association rules of the processed database. To remedy this problem, fuzzy logic was introduced in the association rule extraction process to form a new category of association rules called fuzzy association rules. They convert numerical data into fuzzy data. This transformation maintains the integrity of the information conveyed by the numerical attributes. To extract fuzzy association rules, several algorithms have been introduced. These algorithms can be divided into two categories. The first category includes algorithms based on the extraction of frequent fuzzy itemsets [10], [1], [11], [12], while the second category includes algorithms based on extracting frequent fuzzy closed itemsets [6], [7], [13]. In the following section, we present these two categories of algorithms.

#### 3.1 Algorithms based on frequent fuzzy itemset extraction

In this category, the fuzzy association rule has the following form:

If X is A, then Y is B.

With (X is A) is his premise of the rule and (Y is B) is his conclusion. This rule is noted (X, A)  $\rightarrow$  (B, Y) where  $X = \{x_1 \dots x_p\}$  and  $Y = \{y_1, \dots, y_n\}$  are two disjoint itemsets.  $A = \{a_1 \dots a_p\}$  and  $B = \{b_1, \dots, b_n\}$  are the sets of fuzzy subsets associated with X and Y.

##### Example:

If the age is young and the account balance is small then the loan is moderate. This rule is represented as follows: {Age, account balance} {Young, small}  $\rightarrow$  {Loan} {Moderate}

The algorithms of this category are based on two phases:

- 1) Find all frequent itemsets
- 2) Generate all fuzzy association rules between frequent fuzzy itemsets having a confidence at least equal to *minconf*.

The first algorithms [5], [12], [14] of fuzzy association rules have been proposed to adopt the Apriori algorithm [15] in fuzzy contexts. They focused on reformulating rule validation measures. They offer formulas for support and confidence measures using fuzzy operations and implications.



They adopt the "test and generate" strategy where the algorithm browses the transactional database by level. The principle of these algorithms is to generate the frequent fuzzy itemsets iteratively.

They generate k-itemset (itemset having k items) then determinate the k+1-candidate itemsets by joining the k-itemsets obtained in the previous iteration and preserving only the frequent itemsets whose supports are greater than or equal to the minsup. This step is repeated until there are no candidate itemsets. The extraction of frequent itemsets step is based on the anti-monotonicity constraint. This constraint means that if we have two itemsets I1 and I2 with, I1 included in I2, then the support (I1) > support (I2). Indeed, all the over-itemsets of a non-frequent itemset are not frequent. This constraint reduces the number of candidates and the search space.

These algorithms require a considerable computation time due to an iterative access to the database and they generate huge number of rules, most of them are redundant rules, and often considered irrelevant. The resolution of this problem has been the subject of several studies [10], [11], [16–21]

### 3.2. Algorithms based on frequent closed fuzzy itemset Extraction

In this category, the fuzzy association rule has the following form:

$$\tilde{r}: \tilde{I}1 \Rightarrow \tilde{I}2$$

Where  $\tilde{I}1, \tilde{I}2 \subseteq \tilde{I} = \tilde{I}\{\alpha_1, \alpha_2, \dots, \alpha_p, \alpha_q, \dots, \alpha_n\}$ ,  $\tilde{I}1 = \{\alpha_1, \alpha_2, \dots, \alpha_p\}$ ,  $\tilde{I}2 = \{\alpha_q, \dots, \alpha_n\}$ .  $\tilde{I}1, \tilde{I}2$  are called, respectively, the premise and the conclusion of the fuzzy rule  $r$ . The value  $\alpha_i$ ,  $i = 1, \dots, n$ , is called the local weight of the element.

*Example:*  $R: A^{0.2}, B^{0.3}, C^{0.1} \rightarrow D^{0.8} E^{0.5}$

**If** the attributes A, B and C, respectively, have at least the values 0.2, 0.3 and 0.1 **then** D and E, respectively, have at least the values 0.8 and 0.5.

This category includes the algorithms for extracting frequent fuzzy closed itemsets [6], [7], [13]. An itemset is said to be closed if it does not have any superset with the same support. These algorithms use the fuzzy formal concept analysis FFCA in the process of extracting association rules. Its principle is to extract the set of frequent closed itemsets, from which a subset of rules is generated. This set of generic rules covers the entire extraction context, which ensures the non-loss of information [7]. These algorithms comprise two steps: the extraction of frequent closed itemsets based on the AFC and then the deduction of the generic bases of the association rules.

## 4. RELATED WORK

Since the setting of the algorithm of [5], many algorithms for extraction fuzzy association rules have been proposed. The major problems of these algorithms are their redundancy, their large number and finally, their degree of relevance. Indeed, several works have tried to deal with these problems. We categorized them into three categories: those that use quality measures, those that remove redundant rules, and those that reduce the extraction context and use ontology. In the following, we present these different categories

#### 4.1. Use of quality measures

In order to deal with the problem of the relevance of the fuzzy association rules, some works propose to use quality measures. These measurements help the expert to validate them. In the following, we will list some of them.

Indeed, [12] proposes a method which consists of applying two measures: the factor of certainty and the concept of a very strong rule. The certainty factor (CF) is defined as follows:

$$CF(A \rightarrow C) = \frac{conf(A \rightarrow c) - supp(C)}{1 - supp(C)} \text{ if } conf(A \rightarrow c) > supp(C) \quad (8)$$

$$CF(A \rightarrow C) = \frac{conf(A \rightarrow c) - supp(C)}{supp(C)} \text{ if } conf(A \rightarrow c) < supp(C) \quad (9)$$

The value of the certainty factor is between -1 and 1. It is positive when the dependence between the premise and the consequence is positive. It is negative when the dependence is negative; finally it is zero when the premise and the consequence are independent.

The second measure used in [12] is the new concept of a very strong rule. Indeed, a fuzzy association rule is said to be very strong if the two rules  $A \rightarrow C$  and  $\neg A \rightarrow \neg C$  are valid. According to [12], the extraction of very strong rules probably leads to the generation of relevant knowledge.

In 2006, [22] used the correlation measure to evaluate the interest of a rule. A rule is not presented to the decision maker if its interest is  $< 1$ .

$$Fcorr(<X: A><Y: B>) = \frac{supp(Z: C)}{(supp(X: A) . supp(Y: B))} \quad (10)$$

[16] extended three measures of intensity of classic association rules to use them as part of the fuzzy association rules. These three measures are lift, conviction and leverage. All these measures are based on the assumption of independence.

The lift is a measure of quality that represents the relationship of independence between the premise and the conclusion of the rule. It is the ratio between the observed support and the expected support under the independence hypothesis. The lift is a symmetrical non-implicative measure. It is sensitive to the size of the data: it is a statistical measure. The values of lift  $\in [0; +1]$ .

Conviction also measures independence but between counterexamples. It is the ratio between the number of counterexamples under the assumption of independence and the number of counterexamples observed. It is a non-symmetrical and implicative measure, unlike Lift measurement. However its values  $\in [0; +1]$  as in the case of the lift.

Leverage measures the difference between the observed support and the expected support under the independence assumption.

The author has proved that the simple substitution of the binary medium by the fuzzy support in the classical formulas of these measurements does not give a correct definition and generates erroneous results. In order to propose a correct definition of the measures, the author has defined the expected support and the expected confidence under independence hypothesis as follows

Let  $\otimes$  be the t-norm, X and Y are fuzzy attributes; the expected support is then equal to:

$$\widehat{fsupp}(X \rightarrow Y) = \sum_{i=1}^n \sum_{j=1}^n \frac{X(o_i) \otimes Y(o_j)}{n^2} \quad (11)$$

The expected confidence is equal to

$$\widehat{conf}(X \rightarrow Y) = \frac{fsupp(X \rightarrow Y)}{fsupp(X)} \quad (12)$$

The author has proposed the definition of the three measures as follows:

$$flift(X \rightarrow Y) = \frac{fsupp(X \rightarrow Y)}{\widehat{fsupp}(X \rightarrow Y)} \quad (13)$$

$$flever(X \rightarrow Y) = fsupp(X \rightarrow Y) - \widehat{fsupp}(X \rightarrow Y) \quad (14)$$

$$fconv(X \rightarrow Y) = \frac{\widehat{fsupp}(X \rightarrow \neg Y)}{fsupp(X \rightarrow \neg Y)} \quad (15)$$

## 4.2 Removal redundant rules

In order to reduce the enormous number of extracted rules, other works have proposed to remove the redundant rules. Each offers its definition of redundant fuzzy association rules.

[23] defined the redundant rule as follows: Let A, B, C be three itemsets,  $A \rightarrow B$  and  $A \rightarrow C$  are redundant rules if there is a valid fuzzy association rule  $A \rightarrow B \cup C$ .

[20] is also among the researchers who presented a new fuzzy association rules extraction algorithm that eliminates the rules redundant. The procedure for pruning redundant rules is based on the idea that the confidence value of the rule should increase by increasing the number of elements in the premise.

He has defined the redundant rule as follows: let A, B, and C be three itemsets. A and B are disjoint itemsets and Q contains the subsets of A.

If  $\max_{C \in Q} (\text{conf}(C \rightarrow B)) \geq \text{conf}(A \rightarrow B)$  then  $A \rightarrow B$  is a redundant rule.

The author has also defined the concept of the strong redundant rule.

If  $\min_{C \in Q} (\text{conf}(C \rightarrow B)) \geq \text{conf}(A \rightarrow B)$  then  $A \rightarrow B$  is a very redundant rule

However, according to [24], the proposed algorithm fails because it sometimes eliminates non-redundant rules.

To overcome this problem, [21] propose an improvement of the algorithm presented by [20] by introducing a new notion called notion of equivalence of fuzzy association rules. Its role is to prevent the generation of redundant rules and to prune the redundant itemsets.

The equivalence rules are defined as follows:

Let  $F = \{B_1, B_2, \dots, B_m\}$  be a fuzzy itemset, where  $B$  is a fuzzy item (label) defined on different attributes and  $m$  is the number of elements ( $m > 1$ ) and  $q$  is a threshold equivalence fixed in advance and higher than the predefined value of the minconf.

If  $\text{conf}(U_{i \neq s} B_i \rightarrow B_s) \geq q, \forall s \in \{1, 2, \dots, m\}$  then the rules generated from  $F$  are equivalence rules.

The principle of using the notion of equivalence during pruning is as follows:

Let  $F = \{B_1, B_2, \dots, B_m\}$  be a fuzzy equivalence itemset,  $G = \{B_1, B_2, \dots, B_n\}$  a fuzzy itemset ( $F$  includes  $G$ ), where  $B$  is a fuzzy item (label) defined on different attributes and  $m$  and  $n$  are respectively the number of elements of  $F$  and  $G$  ( $m > 1, n > 1, n > m$ ). Let  $q$  be an equivalence threshold and  $R_F$  and  $R_G$  are association rules generated from  $F$  and  $G$ , respectively:

$$R_F : \bigcup_{i \neq s}^m B_i \rightarrow B_s, \exists s \in \{1, 2, \dots, m\}, R_G : \bigcup_{i \neq s}^n B_i \rightarrow B_s$$

If confidence  $(R_F) > q$ , confidence  $(R_G) > q$  then  $R_G$  rules are redundant rules.

### 4.3. Context Reduction and Use of Ontology

This category includes works that proposed to solve the problem of the number and the relevance of fuzzy association rules by reducing the context and using ontology such as [17]. [17] proposes a method that includes two phases. The first is a preprocessing step. It consists of finding the attributes that have similar behaviors and merging them. Indeed, to measure the similarity of behavior between attributes, the authors use Chi-square test  $X^2$ . The second phase consists to use of the ontology to reduce the candidate itemsets. Indeed, the ontology contains taxonomic relations related to each concept, and the semantic relations between them. To generate frequent itemsets, they examine only the relationships between items related to a concept or items related to different concepts having semantic relations in the ontology.

### 4.4. Discussion

To overcome the problem of relevance of the fuzzy rule, [12], [16], [23], [25], [26] have used a measure to test the validity of extracted rules. This evaluation technique is based on the following principle: following the choices of the measure as well as the threshold of validity by the expert, only the rules having a value higher than or equal to the set threshold are retained. However, several problems can arise. The first problem concerns the arbitrary setting of the threshold which may not cover the desired domain. The second problem is related to the number of extracted rules that can be numerous. In these different cases, the expert finds difficulties during the validation of the rules.

In [17], the authors tried to derive the context by measuring the similarity between the attributes. They build several contingency tables which greatly increases the execution time. This approach

also uses the ontology to reduce the number of candidate rules which requires finding or constructing ontology for each context.

[20], [21] have proposed algorithms to remove redundant rules. The deletion of the rules is based on the deletion of the itemsets during the extraction process. However, these algorithms suffer from some problems namely: the suppression of non-redundant rules [20], the fixation of the value of the parameter  $q$  (equivalence threshold) [21], because any variation of this value can produce different results. In addition, the performance of these two algorithms is not valid because in their experiments the authors used medium-sized test bases. So, we cannot predict how the proposed algorithm will behave in case of large database.

Despite these efforts, the problem of the pertinence and the huge number of fuzzy association rules still persists. In order to resolve these problems, we propose a new validation method able to extract a generic basis of fuzzy association rules and validate them automatically based on fuzzy formal concepts analysis and structural equation model.

## **5. VALIDATION METHOD OF FUZZY ASSOCIATION RULES BASED ON SEM**

In order to validate generated fuzzy association rules and present to the decision maker only useful rules, we propose a new method entitled VMFAR-SEM (Validation Method of Fuzzy Association Rules based on Structural Equation Model). It consists in validating the fuzzy association rules by exploiting the structural equation model. In the following, we present the principle of our method and we illustrate it with an example

### **5.1. Structural Equation Model**

Structural equation model is a powerful, versatile, multi-varied and very general analysis technique used to evaluate the validity of hypotheses with empirical data [27]. The structural equation model offers the flexibility to research and interpret theory and data. It also makes it possible to simultaneously estimate several dependency relationships.

There are two types of variables in this model.

- Manifest variable is a directly collected variable (observed, measured).
- Latent variable is a variable that cannot be directly measured. These variables can be estimated from overt variables.

The structural equation model is decomposed into two sub-models:

- Structural model or internal model is a subset of the complete model including relationships between latent variables.
- Measurement model or external model is a subset of the model

In order to estimate all the relations of the model (relation between the latent variable or relation between latent variable and its indicators), there are two approaches LISREL or PLS. In our work we will use the PLS approach.

The Partial Least Square (PLS) approach is one of the approaches to the structural equation model that comes from an earlier theory called least squares estimation [28]. This theory is based on simple and multiple regressions. Thus, it requires few assumptions and hence its name "soft modeling".

## 5.2. Principle of the proposed method

Our method consists in validating the fuzzy association rule. First we apply our algorithm EFAR-PN [13] to extract generic basis of fuzzy association rule. Then, we apply two steps: The first is to classify rules into groups, according to the items of the premises and conclusions. The second validates the rules by applying the PLS approach on the representative rule of each group. The architecture of our method is shown in Figure 1.

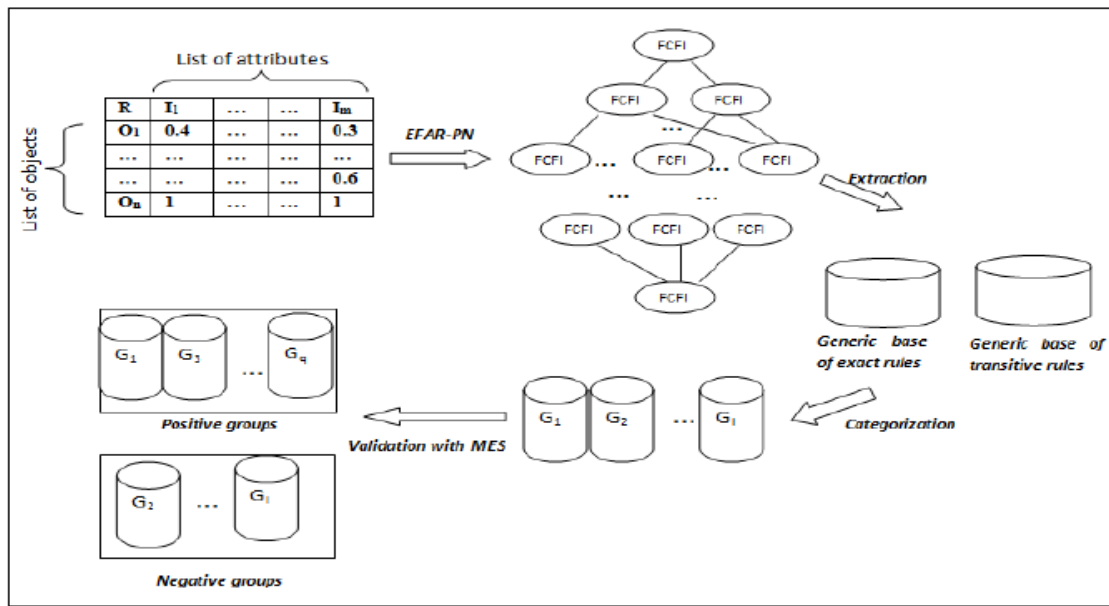


Figure 1. The general architecture of our method

In the following, we explain in detail the principle of each step.

### 5.2.1. Extraction of generic bases of fuzzy association rules

In this step, we apply our algorithm EFAR-PN. This algorithm makes it possible to extract non-redundant rules and without loss of information, while minimizing the execution time. In fact, it extracts frequent fuzzy itemsets without repetitive access to the extraction context and then determines the generic basis of the exact and approximate fuzzy association rules by constructing the iceberg lattice. The construction of the lattice is based on a system of encoding prime numbers. These databases provide the user with a reduced subset of Fuzzy Association Rules (RAFs) covering the entire initial retrieval context and with as much relevant and useful knowledge as possible.

Our EFAR-PN algorithm (Extraction of Fuzzy Association Rules based on the Prime Numbers) has three steps. The first step is to extract the fuzzy minimal generators (GMFFs) by performing a

single access to the fuzzy formal context. The second step consists of building the fuzzy minimal generators lattice. In the third step, the ERAF-NP algorithm deduces the Iceberg lattice of frequent closed fuzzy itemsets (IFFF) and extracts the fuzzy association rules from the lattice of frequent fuzzy minimal generators. (for more details please refer to [13]).

### 5.2.2. Categorization of Fuzzy Association Rules

This step consists of classifying fuzzy association rules into groups. The rules having the same attributes in the premise and the same attributes in the conclusion belong to the same group. Each group has a representative rule. This grouping allows giving a synthetic view of the rules to the user which facilitates their interpretation.

#### For example

Let have following rules:

$$\begin{aligned} R1: & A^{0.5}, B^{0.6}, C^{0.2} \rightarrow D^{0.5}, E^{0.2} \\ R2: & A^{0.6}, C^{0.3} \rightarrow D^{0.6}, F^{0.1} \\ R3: & A^{0.3}, B^{0.2}, C^{0.8} \rightarrow D^{0.7}, E^{0.3} \\ R4: & A^{0.2}, C^{0.7} \rightarrow D^{0.2}, F^{0.7} \end{aligned}$$

Then, we have two groups: G1 contains R1 and R3 and have  $A B C \rightarrow D E$  as representative rule. G2 include two rules R2 and R4 and have  $A C \rightarrow D$  as representative rule.

After classifying the rules into groups, we will apply the validation step.

### 5.2.3. Validation of rules using SEM

In this step, we apply the PLS approach to each representative rule cleared during the previous step. Each rule is considered a model of structural equations that contains two latent variables. The indicators of the first latent variable are the attributes of the premise and the indicators of the second are attributes of the conclusion as shown in Figure 2.

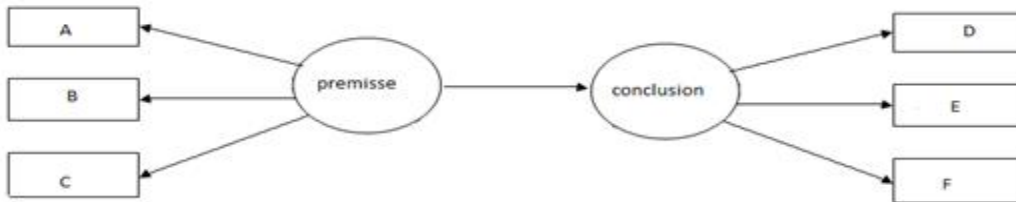


Figure 2. The structural equation model to Measure the  $ABC \rightarrow DEF$

After creating the model, we apply PLS approach to estimate the different coefficients of the model. We first verify the validity of the model. PLS offers several indices measuring the intensity with which the model replicates the database. We will use the coefficient of determination  $R^2$  and  $\alpha$  of cronbach in our method

Then, we interpret the coefficients between the two latent variables (premise and conclusion) and the latent variable and its indicators (items).

The coefficient between the two latent variables is the slop of a simple line regression. This line has the following equation:

$$\text{Conclusion} = a * \text{premise} + b$$

With a is the slope of the line and b is the value at the origin.

If a is positive, then any increase of the premise leads to an increase of the conclusion. A high value of the slope indicates a significant influence of the premise in the conclusion. Indeed, a small change in the premise conducts to a big change in the conclusion.

### 5.3. Illustration of our method

In order to explain the progress of our method, we illustrate its different steps through an example. We will apply our method on the extraction context shown in Table 1.

Table1. Fuzzy extraction context

B	C	E	M
0.4	0.3	0.2	0.3
0.8	1	0.2	0.6
0.8	1	0.5	0.6
1	0.3	1	1

In the following, we apply the steps of our method:

#### First step:

We apply our algorithm EFAR-PN on the extraction context with a minsup=0.25 and minconf=0.5. The result of this step is shown in table 2. We have 23 association rules. The base of exact association rules contains 12 rules while the base of transitive fuzzy association rules contains 11 rules.

Table 2. Bases of exact and transitive fuzzy association rules

Base of exact fuzzy association rules		Base of transitive fuzzy association rules		
Rules	Support	Rules	Support	Confidence
$B^{0.4} \rightarrow C^{0.3}, E^{0.2}, M^{0.3}$	1	$B^{0.4} \rightarrow C^{0.3}, E^{0.2}, M^{0.6}$	0.75	0.75
$C^{0.3} \rightarrow B^{0.4}, E^{0.2}, M^{0.3}$	1	$C^{0.3} \rightarrow B^{0.8}, E^{0.2}, M^{0.6}$	0.75	0.75
$E^{0.2} \rightarrow B^{0.4}, C^{0.3}, M^{0.3}$	1	$E^{0.2} \rightarrow B^{0.8}, C^{0.3}, M^{0.6}$	0.75	0.75
$M^{0.3} \rightarrow B^{0.4}, C^{0.3}, E^{0.2}$	1	$M^{0.3} \rightarrow B^{0.8}, C^{0.3}, E^{0.2}$	0.75	0.75
$B^{0.8} \rightarrow C^{0.3}, E^{0.2}, M^{0.6}$	0.75	$B^{0.8} \rightarrow C^{1.0}, E^{0.2}, M^{0.6}$	0.5	0.6
$M^{0.6} \rightarrow B^{0.8}, C^{0.3}, E^{0.2}$	0.75	$B^{0.8} \rightarrow C^{0.3}, E^{0.5}, M^{0.6}$	0.5	0.6
$C^{1.0} \rightarrow B^{0.8}, E^{0.2}, M^{0.6}$	0.5	$M^{0.6} \rightarrow B^{0.8}, C^{1.0}, E^{0.2}$	0.5	0.6
$E^{0.5} \rightarrow B^{0.8}, C^{0.3}, M^{0.6}$	0.5	$M^{0.6} \rightarrow B^{0.8}, C^{0.3}, E^{0.5}$	0.5	0.6
$B^{1.0} \rightarrow C^{0.3}, E^{1.0}, M^{1.0}$	0.25	$C^{1.0} \rightarrow B^{0.8}, E^{0.5}, M^{0.6}$	0.25	0.5
$E^{1.0} \rightarrow B^{1.0}, C^{0.3}, M^{1.0}$	0.25	$E^{0.5} \rightarrow B^{1.0}, C^{0.3}, M^{1.0}$	0.25	0.5
$M^{1.0} \rightarrow B^{1.0}, C^{0.3}, E^{1.0}$	0.25	$E^{0.5} \rightarrow B^{0.8}, C^{1.0}, M^{0.6}$	0.25	0.5
$C^{1.0}, E^{0.5} \rightarrow B^{0.8}, M^{0.6}$	0.25			



**Second step:**

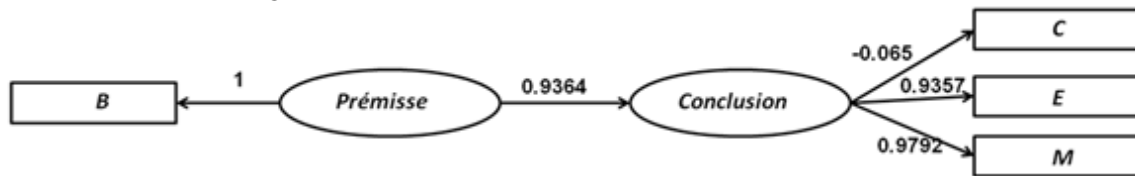
In this step, we divide the extracted fuzzy association rules into groups. The result of applying this step is shown in Table 3. This step allows summarizing the 23 rules in 5 groups.

Table 3. Fuzzy association rule groups

1	Exact fuzzy association rules	Transitive fuzzy association rules	Representative rule
<b>G1</b>	$B^{0.4} \rightarrow C^{0.3}, E^{0.2}, M^{0.3}$ $B^{0.8} \rightarrow C^{0.3}, E^{0.2}, M^{0.6}$ $B^{1.0} \rightarrow C^{0.3}, E^{1.0}, M^{1.0}$	$B^{0.4} \rightarrow C^{0.3}, E^{0.2}, M^{0.6}$ $B^{0.8} \rightarrow C^{1.0}, E^{0.2}, M^{0.6}$ $B^{0.8} \rightarrow C^{0.3}, E^{0.5}, M^{0.6}$	$B \rightarrow C, E, M$
<b>G2</b>	$C^{0.3} \rightarrow B^{0.4}, E^{0.2}, M^{0.3}$ $C^{1.0} \rightarrow B^{0.8}, E^{0.2}, M^{0.6}$	$C^{0.3} \rightarrow B^{0.8}, E^{0.2}, M^{0.6}$ $C^{1.0} \rightarrow B^{0.8}, E^{0.5}, M^{0.6}$	$C \rightarrow B, E, M$
<b>G3</b>	$E^{0.2} \rightarrow B^{0.4}, C^{0.3}, M^{0.3}$ $E^{0.5} \rightarrow B^{0.8}, C^{0.3}, M^{0.6}$ $E^{1.0} \rightarrow B^{1.0}, C^{0.3}, M^{1.0}$	$E^{0.2} \rightarrow B^{0.8}, C^{0.3}, M^{0.6}$ $E^{0.5} \rightarrow B^{1.0}, C^{0.3}, M^{1.0}$ $E^{0.5} \rightarrow B^{0.8}, C^{1.0}, M^{0.6}$	$E \rightarrow B, C, M$
<b>G4</b>	$M^{0.3} \rightarrow B^{0.4}, C^{0.3}, E^{0.2}$ $M^{0.6} \rightarrow B^{0.8}, C^{0.3}, E^{0.2}$ $M^{1.0} \rightarrow B^{1.0}, C^{0.3}, E^{1.0}$	$M^{0.3} \rightarrow B^{0.8}, C^{0.3}, E^{0.2}$ $M^{0.6} \rightarrow B^{0.8}, C^{1.0}, E^{0.2}$ $M^{0.6} \rightarrow B^{0.8}, C^{0.3}, E^{0.5}$	$M \rightarrow B, C, E$
<b>G5</b>	$C^{1.0}, E^{0.5} \rightarrow B^{0.8}, M^{0.6}$		$C, E \rightarrow B, M$

**Third step:**

In this step, we build a model of structural equations for each representative rule. We estimate these models using the approach PLS. We start with the representative rule  $B \rightarrow CEM$ . The result obtained is shown in Figure 3.

Figure 3. Structural Equation Model for  $B \rightarrow CEM$ 

According to the first model of structural equations applied to the representative rule of the first group ( $B \rightarrow CEM$ ), we find that the premise has a strong positive influence on the conclusion (coef = 0.936) and that the increase from B leads to:

- The decrease in C
- The increase of E
- The increase of M

Table 4 shows the result of applying this step on the rules representative of the rest of groups obtained in the first step.

Table 4. The coefficients of the model of the representative rules

Group	Coefficients of the premise indicators	Coefficient between premise and conclusion	Coefficients of the conclusion indicators
G2	1 C	-0.486664263392288	0.669890634808308 <b>B</b> 0.992430570086382 <b>E</b> 0.847801045943832 <b>M</b>
G3	1 <sup>E</sup>	0.945169821893008	0.773122361529185 <b>B</b> -0.43997389929851 <b>C</b> 0.937682683668112 <b>M</b>
G4	1 <b>M</b>	0.954856975574631	0.84089923263202 <b>B</b> -0.284435741070428 <b>C</b> 0.988469944399833 <sup>E</sup>
G5	0.346530943591408 <b>C</b> -0.999259806037586 <b>E</b>	-0.867319793400418	0.979881037935833 <b>B</b> 0.991403087791832 <b>M</b>

We can conclude from table 4 the following remarks:

- According to the second model  $C \rightarrow BEM$ , we notice that the premise has a negative effect on the conclusion (coef = -0.48). Hence the increase of C leads to a decrease in the conclusion set B, E and M.
- According to the third model  $E \rightarrow BCM$ , we synthesize that the premise has a positive and significant influence with a coefficient equal to 0.94. By Therefore, the increase of the premise (increase of E) leads to a increase of the conclusion (increase of B and M and decrease of C).
- According to the fourth model  $M \rightarrow BCE$ , we conclude that the premise has a positive impact on the conclusion and that the increase of one generates the increase of the other. The increase in M leads to the increase of B, the decrease of C and the increase of E.
- According to the fifth model  $CE \rightarrow BM$ , we find that the premise has a negative effect on the conclusion. The increase of C with the decrease of E leads to a decrease of all the conclusion attributes (B and M).

## 6. EXPERIMENTAL STUDY

In this section, we will evaluate our method through using three basics of Fars2008 test, Pendigits and Abalone.

- Abalone<sup>1</sup>: is a base that comes from the UCI Machine Learning Repository. This base represents the physical measurements of the shells. Each is described with 8 variables. This database contains 4177 instances.
- Fars-2008<sup>2</sup>: Data from this database come from the US FARS archive (Fatality Analysis Recording System) which aims at including all accidents in which there has been at least one death. The data concern automobiles where the front passenger seat was occupied, with one observation for each passenger.

<sup>1</sup> <https://archive.ics.uci.edu/ml/datasets.html>

<sup>2</sup> <https://vincentarelbundock.github.io/Rdatasets/datasets.html>

- **Pendigits1**: Pen-Based Recognition of Handwritten Digits Data Set, is a database available in the UCI Machine Learning Repository. This database is produced by collecting 250 examples of 30 writers, written on a pressure sensitive tablet that sent the pen location at fixed time intervals of 100 milliseconds.

The main characteristics of these databases are described in Table 5.

Table 5. Characteristics of the bases of test

Base	Number of objects	Number of attributes
Abalone	4177	8
Fars-2008	64881	24
Pendigits	7494	17

#### First step:

We apply our algorithm on each base of test. Table 6 shows the number of rules extracted in each database with a value of minsup equal to 0.8.

Table 6. Number of extracted rules

Base	Number of exact rules	Number of transitive rules
Pendigits	1988	6366
Abalone	1320	5526
Fars 2008	87	207

We notice that the number of rules can be high even with a high minsup (0.8).

#### Second step:

During the second step, we will divide the extracted rules into groups and identify the representative rule of each group. Table 7 shows the number of groups released for each test basis. We find that this step significantly reduces the number of rules. This step reduces 96% for Pendigits, 87% for Abalone and 77.78% for Fars2008. This reduction makes it easier for the user to explore generated knowledge.

Table 7. The result of the application of the categorization step on the three bases.

Base	Number of fuzzy association rules	Number of groups
Pendigits	8354	321
Abalone	6846	903
Fars2008	294	65

#### Third step:

After determining the different groups, we will apply the second step. The latter is the validation of the rules using PLS. For each group, we construct a structural equation model of its representative rule. The result of this step is shown in table 8.

Table 8. The result of the application of the validation step on the three bases.

Base	Positive groups	Negative groups
Pendigits	265(7933 règles)	56(521 règles)
Abalone	731(5581 règles)	172(1265 règles)
Fars2008	18(76 règles)	47(218 règles)

Positive groups contain rules with a positive PLS coefficient. In these rules, the premise has a positive influence on the conclusion.

Negative groups contain the rules with negative coefficient of degree. In these rules, the premise has a negative influence on the conclusion. We order the representative rules according to the coefficient. The more the coefficient is big and the impact is significant, the more the rule is relevant.

## 7. CONCLUSION

In this article, we have presented our method of validation of fuzzy association rules based on the structural equations model. This method has three steps. The first is to extract generic bases of fuzzy association rules using EFAR-PN algorithm based on fuzzy formal concept analysis. The second step is to classify the rules into groups according to their attributes and to determine a representative rule for each group. This provides a synthetic view of the rules. The third is to construct a model of structural equations from each representative rule. We applied the PLS approach to estimate model coefficients. The PLS coefficient makes it possible to check if the premise has a positive or negative influence on the conclusion. This information can be very useful in many areas. In the future work we plan to create an interactive prototype that integrates the various contributions and that allows the visualization of the rules for the expert to validate them easily. We also plan to test our VMFAR-SEM method on a real world application such as marketing or biology and validate it with a domain expert. For the validation of fuzzy gradual rules, we also intend to apply our VMFAR-SEM method to validate them in a semi-automatic way.

## REFERENCES

- [1] S. Papadimitriou and S. Mavroudi, "The Fuzzy Frequent Pattern Tree," in Proceedings of the 9th WSEAS International Conference on Computers, 2005, pp. 3:1–3:7.
- [2] P. Rajendran and M. Madheswaran, "Novel Fuzzy Association Rule Image Mining Algorithm for Medical Decision Support System," International Journal of Computer Applications, vol. 1, no. 20, pp. 87–94, 2010.
- [3] N. Gupta, N. Mangal, K. Tiwari, and P. Mitra, "Mining quantitative association rules in protein sequences," Data Mining, pp. 273–281, 2006.
- [4] R. Natarajan and B. Shekar, "Interestingness of association rules in data mining: Issues relevant to e-commerce," Sadhana, vol. 30, no. 2–3, pp. 291–309, 2005.
- [5] K. C. Chan and W.-H. Au, "Mining fuzzy association rules," Proceedings of the sixth international conference on Information and knowledge management, pp. 209–215, 1997.

- [6] S. Ben Yahia and A. Jaoua, "Data Mining and Computational Intelligence," J. Kacprzyk, A. Kandel, M. Last, and H. Bunke, Eds. Heidelberg, Germany, Germany: Physica-Verlag GmbH, 2001, pp. 167–190.
- [7] S. Ayouni, "Etude et extraction de regles graduelles floues: définition d'algorithmes efficaces," 2012.
- [8] M. Kryszkiewicz, "Concise representations of association rules," Pattern Detection and Discovery, pp. 92–109, 2002.
- [9] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, "Mining minimal non-redundant association rules using frequent closed itemsets," Computational Logic—CL 2000, pp. 972–986, 2000.
- [10] A. Mangalampalli and V. Pudi, "FPrep: Fuzzy clustering driven efficient automated pre-processing for fuzzy association rule mining.," in FUZZ-IEEE, 2010, pp. 1–8.
- [11] C.-H. Chen, T.-P. Hong, and Y. Li, "Fuzzy association rule mining with type-2 membership functions," Intelligent Information and Database Systems, pp. 128–134, 2015.
- [12] M. Delgado, N. Marín, D. Sánchez, and M. Vila, "Fuzzy association rules: general model and applications," IEEE Transactions on Fuzzy Systems, vol. 11, pp. 214–225, 2003.
- [13] I. Mguiris, H. Amdouni, and M. M. Gammoudi, "An Algorithm for Fuzzy Association Rules Extraction Based on Prime Number Coding," in 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2017, Poznan, Poland, June 21–23, 2017, 2017, pp. 182–184.
- [14] C. M. Kuok, A. Fu, and M. H. Wong, "Mining fuzzy association rules in databases," ACM Sigmod Record, vol. 27, no. 1, pp. 41–46, 1998.
- [15] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, pp. 487–499, 1994.
- [16] M. Burda, "Interest measures for fuzzy association rules based on expectations of independence," Advances in Fuzzy Systems, vol. 2014, p. 2, 2014.
- [17] Z. Farzanyar and M. Kangavari, "Efficient mining of Fuzzy Association Rules from the Pre-Processed Dataset," Computing and Informatics, vol. 31, no. 2, pp. 331–347, 2012.
- [18] J. C.-W. Lin, T.-P. Hong, and T.-C. Lin, "A CMFFP-tree Algorithm to Mine Complete Multiple Fuzzy Frequent Itemsets," Appl. Soft Comput., vol. 28, no. C, pp. 431–439, Mar. 2015.
- [19] R. Prabamanieswari, "Article: A Combined Approach for Mining Fuzzy Frequent Itemset," IJCA Proceedings on International Seminar on Computer Vision 2013, vol. ISCV, pp. 1–5, Jan. 2014.
- [20] T. Watanabe, "Fuzzy association rules mining algorithm based on output specification and redundancy of rules," Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, pp. 283–289, 2011.
- [21] T. Watanabe and R. Fujioka, "Fuzzy association rules mining algorithm based on equivalence redundancy of items," Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, pp. 1960–1965, 2012.

- [22] M. Kaya and R. Alhajj, "Utilizing genetic algorithms to optimize membership functions for fuzzy weighted association rules mining," *Applied Intelligence*, vol. 24, no. 1, pp. 7–15, 2006.
- [23] Y. Gao, J. Ma, and L. Ma, "A new algorithm for mining fuzzy association rules," in *Machine Learning and Cybernetics*, 2004. Proceedings of 2004 International Conference on, 2004, vol. 3, pp. 1635–1640 vol.3.
- [24] A. Roy and R. Chatterjee, "A survey on fuzzy association rule mining methodologies," *IOSR J. Comput. Eng.(IOSR-JCE)*, e-ISSN, pp. 2278–661, 2013.
- [25] M. Kaya, R. Alhajj, A. Arslan, and others, "Efficient automated mining of fuzzy association rules," in *International Conference on Database and Expert Systems Applications*, 2002, pp. 133–142.
- [26] S. Lotfi and M. Sadreddini, "Mining fuzzy association rules using mutual information," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2009, vol. 1.
- [27] W. W. Chin, "The partial least squares approach to structural equation modeling," *Modern methods for business research*, vol. 295, no. 2, pp. 295–336, 1998.
- [28] H. Wold, "Soft modeling: the basic design and some extensions," *Systems under indirect observation*, vol. 2, pp. 589–591, 1982.

## AUTHORS

**Imen Mguiris** received her Master degree in Computer Science at ISIMS-Tunisia in 2010. Now, she prepared her PhD at the Faculty of Sciences of Tunis. Her main research contributions concern: data mining, Fuzzy Association Rules, Formal Concept Analysis (FCA). She is member of Research Laboratory RIADI



**Hamida Amdouni** obtained her Ph.D. and DEA in Computer Science from the Faculty of Science of Tunis respectively in 2014 and 2005. She is currently Assistant Professor at the School of Digital Economy (ESEN), University from Manouba. Her main areas of research are Data Mining, Formal Concepts Analysis (CFA), CRM and Big Data. She is also a member of the SCO-ECRI team at the RIADI Research Laboratory.



**Mohamed Mohsen Gammoudi** is currently a full Professor in Computer Science Department at ISAMM, University of Manouba. He is the responsible of the research Team ECRI of RIADI Laboratory. He obtained his HDR in 2005 at the Faculty of Sciences of Tunis (FST). He received his PhD in 1993 at Sophia Antipolis Laboratory (I3S/CNRS) in the team of Professor Serge Miranda, France. He was hired as a Visiting Professor between 1993 and 1997 at the Federal University of Maranhao, Brazil.



# CLASSIFICATION OF ALZHEIMER USING fMRI DATA AND BRAIN NETWORK

Rishi Yadav, Ankit Gautam, Ravi Bhushan Mishra

Computer Science & Engineering, IIT BHU (Varanasi)

## ABSTRACT

*Since the mid of 1990s, functional connectivity study using fMRI (fcMRI) has drawn increasing attention of neuroscientists and computer scientists, since it opens a new window to explore functional network of human brain with relatively high resolution. BOLD technique provides almost accurate state of brain. Past researches prove that neuro diseases damage the brain network interaction, protein- protein interaction and gene-gene interaction. A number of neurological research paper also analyse the relationship among damaged part. By computational method especially machine learning technique we can show such classifications. In this paper we used OASIS fMRI dataset affected with Alzheimer's disease and normal patient's dataset. After proper processing the fMRI data we use the processed data to form classifier models using SVM (Support Vector Machine), KNN (K- nearest neighbour) & Naïve Bayes. We also compare the accuracy of our proposed method with existing methods. In future, we will other combinations of methods for better accuracy.*

## KEYWORDS

*Brain Network, Dementia, Alzheimer's disease, SVM, KNN, Fmri*

## 1. INTRODUCTION

Alzheimer's disease (AD), a progressive, irreversible, neurodegenerative disorder and multifaceted disease, occurs most frequently in older age. Alzheimer's disease slowly destroys brain cells causing loss of memory, thinking skill, behaviour and learning and ultimately ability to perform simple tasks. AD is the only disease among top 10 causes of death among Americans, which cannot be prevented, slowed and cured. Early diagnosis of Alzheimer's disease is very costly, time-taken and careful medical assessment of physical history, parent's health analysis, physical and neurobiological exam and Mini Mental State Examination (MMSE) etc.[13]. Therefore, development in the field of automatic computational Alzheimer's Detection is economically desirable.

Neuronal dysfunction in AD mainly causes due to failure of functional integration and damage in connectivity network of brain[6]. Brain is the centre of human nervous system. Brain is made of more than 100 billion of nerves that communicate in trillions of connections called synapses. It contains spatially distributed but functionally connected regions that continuously share stimulus and responses to each other. In past three decades, a rich study on functional and structural neuroimaging have provided a plenty of knowledge about role, functions and connectivity pattern

of brain regions. A number of computational and biological tools have also developed to collect this information. [5][8][10]. As one's age passes, the brain cells develop plaques and tangles. They first developed in the areas involved in memory and then spread out to other parts of the brain. These tangles and plaques disable or block the communication among the nerve cells and distress its function and eventually the cells will die [8].

For the early detection of Alzheimer's disease, the analysis of neuroimaging data has achieved much attraction recent years. Some neuroimaging modalities or biomarker exists in order to study and detect AD and other brain related diseases are positron emission tomography (PET), magnetic resonance imaging (MRI), computed tomography (CT), study of cerebral metabolism with fluoro-deoxy-d-glucose (FDG) etc. [7]. The past two decades have witnessed the increasing use of resting state functional magnetic resonance imaging (rs-fMRI) as a tool for mapping human brain network. The word 'rest' refers to a constant condition without imposed stimulus i.e. no task performance during acquisition of fMRI data. The resting brain activity is measured through observing the changes in oxygen and blood flow in the brain parts and among the brain parts, which creates a signal referred to blood-oxygen-level dependent (BOLD) signal that can be measured using functional magnetic resonance imaging technique.

The BOLD signals represents low-frequency spontaneous fluctuations of oxygen and blood flow in brain network regions. Initially it had been thought to be a noise. In **Biswal et al., [14]** demonstrated that that these low frequencies (0.1 to 0.01 Hz) in BOLD signal are highly correlated among the brain regions. Later the low frequency fluctuations were shown to be of neural origin and specific to grey matter across the hemispheres in the bilateral motor cortices [17]. Thus the correlation structure of these rs-fMRI can be used to determine the network within the brain.

**Zhang et al., [6]** investigated the functional connectivity in the resting brain network by comparing the samples of healthy volunteer and patient affected with different levels of Alzheimer disease. To detect the alteration in brain network in general and posterior cingulate cortex (PCC) in particular, resting state functional magnetic resonance imaging used by the comparing of fMRI dataset. Data acquisition was performed following all the ethical and scientific protocols prescribed by Diagnostic and Statistical Manual of Mental Disorders. The study observed that the set of regions like the hippocampus, the inferior temporal cortex, the visual cortices and especially the presumes and cuneus, the medial prefrontal cortex having high degree of dissociated functional connectivity with PCC in all AD patients. Study also observed that the degree of connectivity level also intensified as the stage of AD progression increased from mild to moderate and then severe AD. Thus study concluded that alteration in functional connectivity in brain network might play a role in early diagnosis of AD.

**Prasad et al.,[8]** compared a variety of anatomic connectivity measures, including several novel ones like global efficiency, transitivity, path length, modularity, local efficiency, optimal community structure, eigenvector centrality etc. that may help in distinguishing Alzheimer's disease(AD) patients. They evaluated two kinds of connectivity measures. The first evaluated measures from whole-brain tractography connectivity matrices and second one studied additional network measures based on a novel flow-based measure of connectivity matrices. The study evaluated the measures' ability to discriminate disease by using 10-fold cross-validated classifier that were repeated 30 times and found the highest accuracy of 78.2%.



**Biju K Sa et al.,[1]** provided the software solution of for detecting the Alzheimer's disease and brain abnormalities as well and produced a 3D representation of MRI slices. The study collected and analyzed various parameters from MRI dataset like cortex area, grey matter volume, white matter volume, and cavity area and brain density. They used grey to white matter ratio for determining if person is affected by Alzheimer's disease. MRI slices undergo with different processes like segmentation, calculation of density and volume of brain parts, de-noising and 3D construction. Finally the study concluded that the grey matter to white matter volume ratio will be more for the person having brain abnormalities. This study included brain abnormalities such as tumour and internal blood clotting etc. too.

**Sarraf et al.,[2]** used convolutional neural network (CNN) to classify Alzheimer brain and normal healthy brain. They used MRI slices of Alzheimer patient and normal healthy brain from ADNI databank and pre-processed the image using the standard modules of FMRIB library v5.0. Images were labelled for binary classification of Alzheimer's Vs normal dataset and these labelled images were converted to Imdb storage databases for higher throughput to be fed into Deep learning platform. The study used CNN and famous architecture LeNet-5 and successfully classified fMRI data of Alzheimer's patient from normal controls. The accuracy of test data on training data reached 96.85 %, which was trained and tested with huge number of images. This study also showed a novel path to use more complicated computational architecture and classifiers to increase the efficient and effective pre-clinical diagnosis of AD.

**F. Previtali et al.,[7]**Extracted the key features from magnetic resonance imaging which are most suitable for classification task. For extraction of key features the study proposed a novel feature extraction technique that is based on recent computer vision method, called Oriented FAST and Rotated BRIEF. They used the state-of-the-art approaches on two medical established databank ADNI and OASIS. The extracted features are processed with the combination of two new metrics i.e., their spatial position and their distribution around the patient's brain, and given as input to a function-based classifier, Support Vector Machines (SVM). The study obtained a classification accuracy and sensitivity of 77% and specificity of 79% when dealing with four classes using OASIS dataset.

**Behashti et al.,[5]** presented an automatic computer-aided diagnosis(CAD) system for detection of Alzheimer's disease using structural MRI dataset from ADNI. The proposed method consists four stages. First, local and global difference in grey matter (GM) of AD patient and GM of normal healthy patient. Second, the voxel intensity values of the VOIs are extracted as raw feature. Third, the raw features are ranked by using seven-feature ranking methods. The feature with higher score value are more discriminative. Fourth, the Support Vector Machine (SVM) classifier used for classification. The classification accuracy of this proposed method for diagnosis of AD is up to 92.48% using sMRI data from state-of-the art databank ADNI.

We used state-of-the-art resting state functional magnetic resonance imaging (rs-fMRI) dataset from OASIS (Open Access Series of Imaging Studies). Dataset consists MRI slices of 416 subjects. We used MATLAB for image pre-processing steps- Realignment, Normalization, Segmentation, Outlier Detection, De-noising and smoothing. We got features of fMRI image to draw and analyse the distorted brain network of Alzheimer's Patient and the difference from the normal patient's brain network. We have deployed efficient intelligent computing methods in MATLAB for classification and diagnosis of Alzheimer's disease and generate a comparative study of accuracy obtained by these methods. We have used classifiers respectively Support Vector Machine (SVM), K-Nearest Neighbour (KNN) classifier and Naïve based Classifier. To

further increase the accuracy, we used most relevant features using feature reduction principal component analysis (PCA) method and feature selection Maximum relevance minimum redundancy (MRMR) method. The proposed method achieved high accuracy via using 10-fold cross validation technique.

The comparative analysis of literature can be summarized by the below table.

Table-1: Comparative analysis of literature survey

Paper	Methods	Results
Resting Brain Connectivity: Changes during the progress of Alzheimer Disease	Temporal correlation method used to obtain PCC connectivity maps.	Functional connectivity between the PCC and a set of regions
Structural MRI-based detection of Alzheimer's disease using feature ranking and classification error	CAD system	Classification accuracy of the proposed automatic system for the diagnosis of AD is up to 92.48% using the sMRI data
Classification of Alzheimer Diagnosis from ADNI Plasma Biomarker Data	five conventional classification algorithms: libSVM	The accuracy is 86% for the ensemble.
Classification of Alzheimer's Disease Using fMRI Data and Deep Learning Convolutional Neural Networks	Deep learning, CNN	Accuracy of 82%
A novel method and software for automatically classifying Alzheimer's disease patients by magnetic resonance imaging analysis	Computer vision method, called <i>Oriented FAST and Rotated BRIEF</i>	Accuracy of 79%
Brain connectivity and novel network measures for Alzheimer's disease classification	SVM	Brain network analysis of degree and betweenness parameter

The paper's organization is as follows, Apart from introduction, Section II deals with problem description, section III deals with proposed methods including data acquisition, data pre-processing, proposed algorithm, section IV deals with results and discussion and section V represents conclusion.

## 2. PROBLEM DESCRIPTION

Early diagnosis of Alzheimer's disease (AD) via medical assessment is very costly and time consuming. The purpose of this paper is to propose a computational model which can differentiate the Alzheimer's disease patient and normal patient dataset. We have resting state

functional magnetic resonance imaging (rs-fMRI) dataset from OASIS as input for computational models.

### **3. PROPOSED METHOD**

#### **3.1 Data Acquisition**

Resting state functional magnetic resonance images (rs-fMRI) were obtained from state-of-the-art Open Access Series of Imaging Studies (OASIS) from (<http://www.oasis-brains.org/>). OASIS provide brain imaging data that are freely available for data analysis. The dataset consists of cross sectional images of 416 subjects having age 18 to 96 and covers all stages of AD. Additionally, images from a subsequent scan session after a short delay (less than 90 days) are also included as a means of assessing acquisition reliability, for 20 of the non-demented subjects. OASIS dataset also contains demographic, clinical, and derived anatomic measures located in the spreadsheets files (oasis\_cross-sectional.xls and oasis\_cross-sectional.csv). We used .nii format MRIs of all the subjects as our input dataset.

#### **3.2 Data Pre-processing**

Each subject has multiple slices in their MRI. The total number of slices in each subject's MRI is equal to 198. We applied functional Realignment for subject motion estimation and correction. After that, we applied translation of functional centre to (0, 0, 0). Functional Slice-timing correction. MRIs are usually measured using 2D imaging methods repeatedly and this technique can result in a temporal offset difference between slices. We have applied slice-timing correction to mitigate the offset.

MRIs often contain unusual and dissimilar observations also called as outliers. This can happen for numerous reasons, for example, data acquisition, pre-processing artefacts, resulting from variance between large natural inter-subjects. But homogeneous samples are assumed in all of the statistical procedure. We have already centred the data using translation and we can use outlier detection to remove the dissimilarities.

To achieve simultaneous Grey/White/CSF segmentation, we applied functional direct segmentation and MNI Normalisation.

De-noising is also done using linear regression and band-pass filtering to remove physiological, unwanted motion, and other artefactual effects from the BOLD signal before computing connectivity measures. By default the system will start with three different sources of possible confounders: 1) BOLD signal from the white matter and CSF masks (5 dimensions each); 2) any previously-defined within-subject covariate (realignment and scrubbing parameters); and 3) the main condition effects (condition blocks convolved with hrf). For each of the selected possible confounds you may change the number of dimensions (specifying how many temporal components are being used), and the derivatives order (specifying how many successive orders of temporal derivatives are included in the model). For example, the realignment confound (derived from the estimated subject motion parameters) is defined by default by 6 dimensions. You can change the derivative order to 1 indicating that in addition the first-order temporal derivative of the motion parameters should also be used as covariates. Similarly, the White Matter confound is

defined by default by 5 dimensions and 0 derivative order (indicating that 5 PCA temporal components are being used, with not additional temporal derivative terms).

### 3.3 Computing Methods

After the data pre-processing step, we use measures of efficiency, centrality, and cost/degree, associated with an ROI-to-ROI connectivity network built after second level analysis. After this, we use some popular machine learning models to build our classifier models to discriminate between Alzheimer's and non-Alzheimer's patients. We have 165 regions on brain network for each patient. We use degrees of each region node as our parameters for training the classifiers.

#### 3.3.1 Naive Bayes

The naive Bayes classifier applies to learning tasks where each instance  $x$  is described by a conjunction of attribute values and where the target function  $f(x)$  can take on any value from some finite set  $V$ . A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values  $(a_1, a_2 \dots a_n, a_1, a_2 \dots a_n)$ . The learner is asked to predict the target value, or classification, for this new instance.

#### 3.3.2 Support Vector Machine (SVM)

Support vector machine is depends on pre-processing the data to higher dimension patterns rather than original feature space. The support vectors are the sample that are highly difficult for classification and the also define the hyperplane which should be optimally separated. The SVM finds the best classification function with largest margin of hyperplane separation between the two classes.

$$L_p = \frac{1}{2} ||\vec{w}'|| - \sum_{i=1}^t a_i y_i (\vec{w}' \cdot \vec{x}_i + b) + \sum_{i=1}^t a_i$$

Where  $L_p$  is called the Lagrange,  $t$  is the number of training samples, and  $a_i$  are the Lagrange multipliers,  $b$  is a constant and  $w$  is the vector that defines the hyperplane.

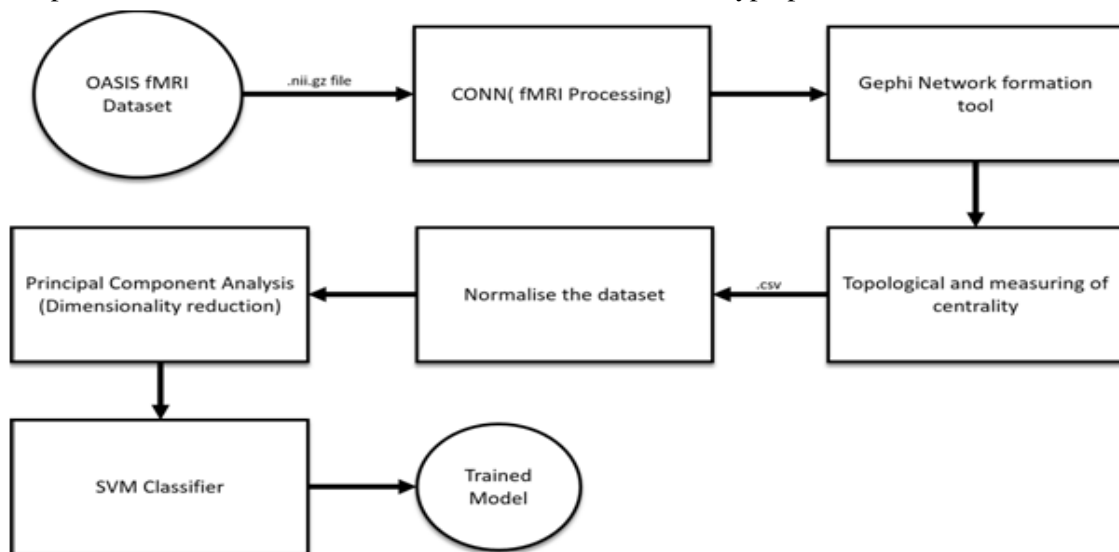


Figure 1. Flowchart of proposed model

### 3.3.3 K- Nearest Neighbour Classifier (KNN)

K-Nearest Neighbour is one of the most basic yet essential classification algorithms in Machine Learning. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM which assume a Gaussian distribution of the given data). In KNN we are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

When a unknown tuple, a k-nearest neighbour classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. The k training tuples are the k nearest neighbour of the unknown tuple.

### 3.3.4 Principle Component Analysis (PCA) feature dimensionality reduction Method

PCA is unsupervised feature transformation and linear projection method as vector projection proceeds without any knowledge of their class labels and reveals hidden information by maximizing the variance. An N-dimensional weighted feature  $x_i (i = 1, 2, 3, \dots, m, N < m)$  was projected on the Eigenvectors of its covariance matrix and transformed matrix (v), the features of Eigenvalues greater than '1' is selected and tends to form a subset of 5 uncorrelated features. By PCA dimensionality is reduced. Few PCs can capture variances of the data. PCs are uncorrelated and ordered. We expect that cluster structure in the original dataset can be extracted from first few PCs.

X represents the original data matrix;  $Y = (y_1, \dots, y_n), y_i = x_i - \bar{x}$ , represents the centred data matrix where  $\bar{x} = \sum_i x_i / n$ . The covariance matrix (ignoring the factor 1/n) is  $\sum_i (x_i - \bar{x})(x_i - \bar{x})^T = Y Y^T$ . Principal directions  $u_k$  and principal components  $v_k$  are eigenvectors satisfying  $Y Y^T u_k = \lambda_k u_k, Y Y^T v_k = \lambda_k v_k, v_k = Y^T u_k / \lambda_k^{1/2}$

### 3.3.5 MRMR Features selection method

Minimal Redundancy Maximum Relevance is a feature selection method. Maximum Relevance means that selecting the features which has highest relevance with target class, based on mutual information, f-test And minimal redundancy means that selected features are correlated with each other and that features are cover narrow regions in the space.

Maximize Relevance:

$$\max V_I, V_I = \frac{1}{|S|} \sum_{i \in S} I(h, i)$$

Where s is the set of features, I (i, j) is mutual information between feature i and j.

Minimal Redundancy:

$$\min W_I, W_I = \frac{1}{|S|^2} \sum_{i, j \in S} I(i, j)$$

### 3.4 Pseudo-code for the proposed model

/\* Performance are measured by k-fold Cross Validation, where k=10 \*/

Begin:

- fMRI processing steps- Realignment, Normalization, Segmentation, Outlier Detection, De-noising and smoothing -> Brain Network Topological and centrality measurement
- For i=1 to k:
  - k-1 subset -> Training set;
  - Remaining subset -> Testing set;
  - Rank features technique, mRMR;
  - Dimension reduction technique, PCA
  - Train SVM, KNN and Naïve bayes classifier on reduced space training dataset using different size of feature subset;
  - Test the trained SVM, KNN and Naïve bayes models on Testing set;
- EndFor;
- Calculate the average classification accuracy of SVM, KNN and Naïve bayes over  $i^{\text{th}}$  testing set;

End.

## 4. RESULTS

As per workflow diagram of proposed model, first we generate brain graph network of more concerned Region of Interest (ROI) of brain related to Alzheimer disease. We used MATLAB for image pre-processing steps- Realignment, Normalization, Segmentation, Outlier Detection, De-noising and smoothing. We got features of fMRI image to draw and analyze the distorted brain network of Alzheimer's Patient and the difference from the normal patient's brain network. Now, using gephi tool, we generate .csv file of topological measures and centrality. As per paper [3], degree and betweenness are two important parameters for classifier model. But in our method we used other parameters average path length, clustering coefficient, cost, local and global efficiency too with degree and betweenness. Then we used PCA for dimensionality reduction and MRMR for relevant feature selection. In compare to paper [3], we got increased accuracy 95% using SVM, 95% using KNN and 90% from Naïve bayes classifier.

Table-2: Sensitivity & Specificity

Method	True Positive	False Positive	True Negative	False Negative
SVM	10	1	9	0
Naive Bayesian	10	2	8	0
k-nearest neighbour	10	1	9	0

## SVM

Initially the features that are extracted from the MRI image are used to train the SVM classifier. In training phase, the model is built while in testing phase, the model is validated. There is only one type of classification performed in which the features are fed to SVM which creates models numbered -1 for Alzheimer's disease and +1 for normal person.

In the testing period, all of the features are input into SVM and distance between each of the vector and hyperplane is calculated. We achieved a true positive rate of 100 % and false positive rate of 10% and the accuracy of our model is 95%.

## Naive Bayesian

Naive Bayesian depends upon the selection of training data for increasing the accuracy of the classifier. Partitioned vectors of processed MRI's features are used to test and train the Naive Bayesian classifier. This also uses the same class models as SVM numbered +1 and -1 for normal person and Alzheimer's disease respectively.

In the testing phase, 20 vectors are validated out of which we received a true positive rate of 100% and a false positive rate of 20% which gives a total accuracy of 90%.

## K-nearest Neighbour

In the testing phase, 20 vectors are validated out of which we received a true positive rate of 100% and a false positive rate of 20 % which gives a total accuracy of 90%.

In compare to other methods proposed by authors [3],[4],[14],[15], we get more accurate classifier model. For the relative comparison refer to table-1 and literature survey.

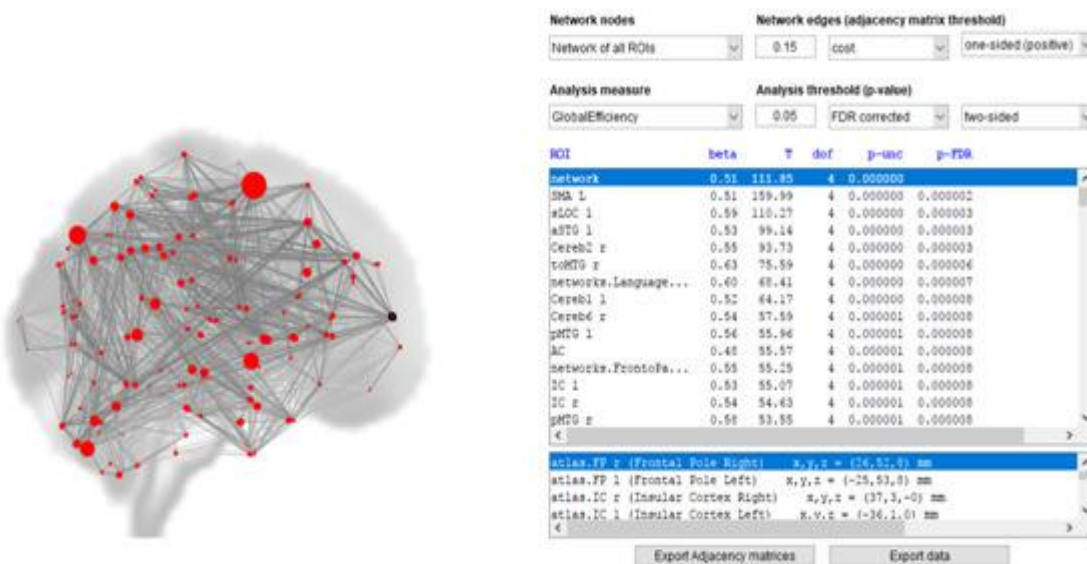


Figure 2. Brain network generated from OASIS AD data

## 5. DISCUSSION & CONCLUSION

Functional magnetic resonance imaging or functional MRI (fMRI) measures brain activity by detecting changes associated with blood flow. Using BOLD (Blood Oxygen level dependent) technique accurate brain data can be captured as fMRI image. We use state-of-the art database OASIS for this purpose. By proper processing of fMRI images and pre-processing approaches we get brain connectivity network. Alzheimer or any other neurodegenerative diseases can be captured by analysing this brain connectivity network. . The grey matter activity of Alzheimer disease patient is not as high as a non-Alzheimer patient. The grey matter stands for oxygen level in the brain. It is intuitive that oxygen level in Alzheimer's disease patient is lower than a normal brain. Oxygen level also shows the amount of activity being done in the brain, so a higher oxygen level brain means active brain. In our proposed method, finally using SVM, KNN and Naïve bayes, we get success in making classifier model with quite good accuracy of 95%, 95% and 90%. In our proposed method we focus on brain connectivity disturbance due to Alzheimer's disease. There are other areas also, which got affected by neuro diseases like blood cells, protein-protein interaction, gene-gene interaction etc. We will try to all incorporate all effective regions during design of classifier model. Other than organic areas like brain network, blood cells, protein and genes, non-organic data like Demographic, clinical, and derived anatomic measures can also be used for classifier model. In this proposed method we used OASIS fMRI dataset of Alzheimer's disease and used SVM classifier model to classify neuro disease. In future we will use other dataset like ADNI and composition of other classifier models like libSVM(an ensemble of five conventional algorithms) as described in paper[14].

## REFERENCES

- [1] Biju, K. S., S. S. Alfa, Kavya Lal, Alvia Antony, and M. KurupAkhil. "Alzheimer's Detection Based on Segmentation of MRI Image." *Procedia Computer Science* 115 (2017): 474-481.
- [2] Sarraf, Saman, and GhassemTofighi. "Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks." *arXiv preprint arXiv:1603.08631* (2016).
- [3] Goñi, Joaquín, Francisco J. Esteban, Nieves Vélez de Mendizábal, Jorge Sepulcre, Sergio Ardanza-Trevijano, Ion Agirrezabal, and Pablo Villoslada. "A computational analysis of protein-protein interaction networks in neurodegenerative diseases." *BMC systems biology* 2, no. 1 (2008): 52.
- [4] Zhao, Jinying, Yun Zhu, Jingyun Yang, Lin Li, Hao Wu, Philip L. De Jager, Peng Jin, and David A. Bennett. "A genome-wide profiling of brain DNA hydroxymethylation in Alzheimer's disease." *Alzheimer's & dementia: the journal of the Alzheimer's Association* 13, no. 6 (2017): 674-688.
- [5] Beheshti, Iman, Hasan Demirel, FarnazFarokhian, Chunlan Yang, Hiroshi Matsuda, and Alzheimer's Disease Neuroimaging Initiative. "Structural MRI-based detection of Alzheimer's disease using feature ranking and classification error." *Computer methods and programs in biomedicine* 137 (2016): 177-193.
- [6] Zhang, Hong-Ying, Shi-Jie Wang, Bin Liu, Zhan-Long Ma, Ming Yang, Zhi-Jun Zhang, and Gao-Jun Teng. "Resting brain connectivity: changes during the progress of Alzheimer disease." *Radiology* 256, no. 2 (2010): 598-606.
- [7] Previtali, Fabio, Paola Bertolazzi, Giovanni Felici, and Emanuel Weitschek. "A novel method and software for automatically classifying Alzheimer's disease patients by magnetic resonance imaging analysis." *Computer methods and programs in biomedicine* 143 (2017): 89-95.



- [8] Prasad, Gautam, Shantanu H. Joshi, Talia M. Nir, Arthur W. Toga, and Paul M. Thompson. "Brain connectivity and novel network measures for Alzheimer's disease classification." *Neurobiology of aging* 36 (2015): S121-S131.
- [9] Wang, Hong, Wenwen Chang, and Chi Zhang. "Functional brain network and multichannel analysis for the P300-based brain computer interface system of lying detection." *Expert Systems with Applications* 53 (2016): 117-128.
- [10] Li, Kaiming, Lei Guo, JingxinNie, Gang Li, and Tianming Liu. "Review of methods for functional brain connectivity detection using fMRI." *Computerized Medical Imaging and Graphics* 33, no. 2 (2009): 131-139.
- [11] Sheline, Yvette I., and Marcus E. Raichle. "Resting state functional connectivity in preclinical Alzheimer's disease." *Biological psychiatry* 74, no. 5 (2013): 340-347.
- [12] Plant, Claudia, Christian Sorg, Valentin Riedl, and AfraWohlschläger. "Homogeneity-based feature extraction for classification of early-stage alzheimer's disease from functional magnetic resonance images." In *Proceedings of the 2011 workshop on Data mining for medicine and healthcare*, pp. 33-41. ACM, 2011.
- [13] Mo, Jue, Sana Siddiqui, Stuart Maudsley, Huey Cheung, Bronwen Martin, and Calvin A. Johnson. "Classification of Alzheimer Diagnosis from ADNI Plasma Biomarker Data." In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, p. 569. ACM, 2013.
- [14] Biswal, Bharat, F. ZerrinYetkin, Victor M. Haughton, and James S. Hyde. "Functional connectivity in the motor cortex of resting human brain using echo-planar mri." *Magnetic resonance in medicine* 34, no. 4 (1995): 537-541.
- [15] Li, Shi-Jiang, Bharat Biswal, Zhu Li, Robert Risinger, Charles Rainey, Jung-Ki Cho, Betty Jo Salmeron, and Elliot A. Stein. "Cocaine administration decreases functional connectivity in human primary visual and motor cortex as detected by functional MRI." *Magnetic Resonance in Medicine* 43, no. 1 (2000): 45-51.

## AUTHORS

**Rishi Yadav** – Final year Integrated dual degree (Btech + Mtech) student in Computer Science and Engineering at Indian Institute of Technology, BHU, Varanasi, India.

Email- rishi.yadav.cse13@iitbhu.ac.in

**Ankit Gautam** - Final year Integrated dual degree (Btech + Mtech) student in Computer Science and Engineering at Indian Institute of Technology, BHU, Varanasi, India.

Email- ankit.gautam.cse13@iitbhu.ac.in

**Ravi Bhushan Mishra**- Professor in Computer Science and Engineering at Indian Institute of Technology, BHU, Varanasi, India.

Email- mishraravi.cse@itbhu.ac.in

*INTENTIONAL BLANK*

# AUTOMATED PENETRATION TESTING: AN OVERVIEW

Farah Abu-Dabaseh and Esraa Alshammari

Department of Computer Science  
Princess Sumaya University for Technology, Amman, Jordan

## **ABSTRACT**

*The using of information technology resources is rapidly increasing in organizations, businesses, and even governments, that led to arise various attacks, and vulnerabilities in the field. All resources make it a must to do frequently a penetration test (PT) for the environment and see what can the attacker gain and what is the current environment's vulnerabilities. This paper reviews some of the automated penetration testing techniques and presents its enhancement over the traditional manual approaches. To the best of our knowledge, it is the first research that takes into consideration the concept of penetration testing and the standards in the area. This research tackles the comparison between the manual and automated penetration testing, the main tools used in penetration testing. Additionally, compares between some methodologies used to build an automated penetration testing platform.*

## **KEYWORDS**

*Penetration test, Automation, Exploitation, Ethical hacker, Penetration testing standards.*

## **1. INTRODUCTION**

Penetration testing is used to check the exploitations and the vulnerability of the organization's system and help the developers to build a protected system that meets the needs. It's very important to any organization and company to protect their data and information from outside attackers and keep monitoring to the prioritize the severity of the security issues. Determining the priorities can help the developers to determine the needed devices in the allocation of the budget for security issues. Additionally, can be used to find the financial loss expected and risks if the attackers achieve their goals and exploited the system and how to mitigate that. The data generated from the test considered confidential and private data because it shows approximately all the holes in the system and how they could be exploited. [1]

PT can be done by attacking the system similar to the action of the outside attackers and find out what can be obtained [2]. The attack might not be as easy as exploiting one vulnerability, many vulnerabilities may be used to achieve the goal by making a sequence of attack chain (Multi-step attack) [3]. It's also considered as a risk assessment and can be used to check the network safety. When penetration test is done, the roles of engagement for that test should be set also, to set the goals and the methodology of the test.

Penetration tests companies can be classified into three different types: gray hat, black hat, and white hat. In the white hat, the tester is an ethical hacker that respects the rules of the organization and the employees can help to perform the testing. While the black hat is mainly used to find how the employees interact with the undesired attack, in this approach the administrators are only the ones who know the test is underway. Moreover, we can do a Gray hat which is a combined approach to the previous types into a custom test plan [4].

Penetration testing should be considered as a standard frequent process within the security roadmap. Traditionally, the organizations used to perform the penetration testing only when they have a product release or a major upgrade. [5]

However, it's suitable to perform the test in these situations:

- New installed software
- Applied system upgrades
- User policy modification
- Applied Security patches
- New infrastructure is added

Although it's important to have a penetration testing in the organization, it's hard to implement too. Since it should include a security expert with the capability to do such a complex job. That could be an overhead on the organization and could waste time and money without the desired result in the case that the security team wasn't as professional as they must. So, the automated approach has seen the light; done by an expert security team in the field.

The contribution of this research will consider the standards of penetration testing, the tools used for each phase in the penetration test, the comparison between the automated and manual approaches and the comparison between some of the current approaches for the automated penetration test.

The rest of this research is as follows: in section two the Penetration testing standards will presents. While the comparison between manual and automated techniques in penetration testing are provided in section three. Section four shows an overview of the current automated penetration testing. Finally, the conclusion and future works are presented in section five.

## **2. PENTRATON TESTING STANDARDS**

Standards for penetration testing aimed to provide a basic outline and definition of the penetration testing. Also to give an outline of the steps used for it, many standards are out there having various pros and cons [6]. Choosing one of them should be based on the goal of having the test.

There are currently various standards that could be followed, such as ISAAF (Information Systems Security Assessment Framework), the OSSTMM (Open-Source Security Testing Methodology Manual), the NIST SP 800-115, and the PTES (the Penetration Testing Execution Standard), the OISSG (Open Information Systems Security Group) [6].

OSSTMM v3 covers the whole parts of the penetration test and have three classes of attacks: Communications Security, Spectrum Security and Physical Security. This standard was published in 2010 and is very mature since the first version which was released in 2000.[6]

On the other hand NIST (SP800-115) standard provides guidelines for planning and conducting information security testing and assessments. Additionally, to analyze the findings and developing mitigation strategies. It's not purposed to give an overall testing or assessment program but to give an overview of the key elements in both security testing and assessment with assurance on specific techniques showing their benefits and limitations. in addition to that, it also gives recommendations and reports for their use.

On the other hand NIST (SP800-115) standard provides guidelines for planning and conducting information security testing and assessments. Additionally, to analyze the findings and developing mitigation strategies. It's not purposed to give an overall testing or assessment program but to give an overview of the key elements in both security testing and assessment with assurance on specific techniques showing their benefits and limitations. in addition to that, it also gives recommendations and reports for their use.

As (SP800-115) NIST standard, the penetration testing process can be divided into the following four processes shown in (Figure.1) [7]:



Figure.1: The Phases of Penetration Testing (NIST) Standard

The third standard in penetration testing is ISSAF methodology which aimed to help the administrator to evaluate your application, system and network controls. It consists of three phases and nine steps [8]. As shown in Figure 2:

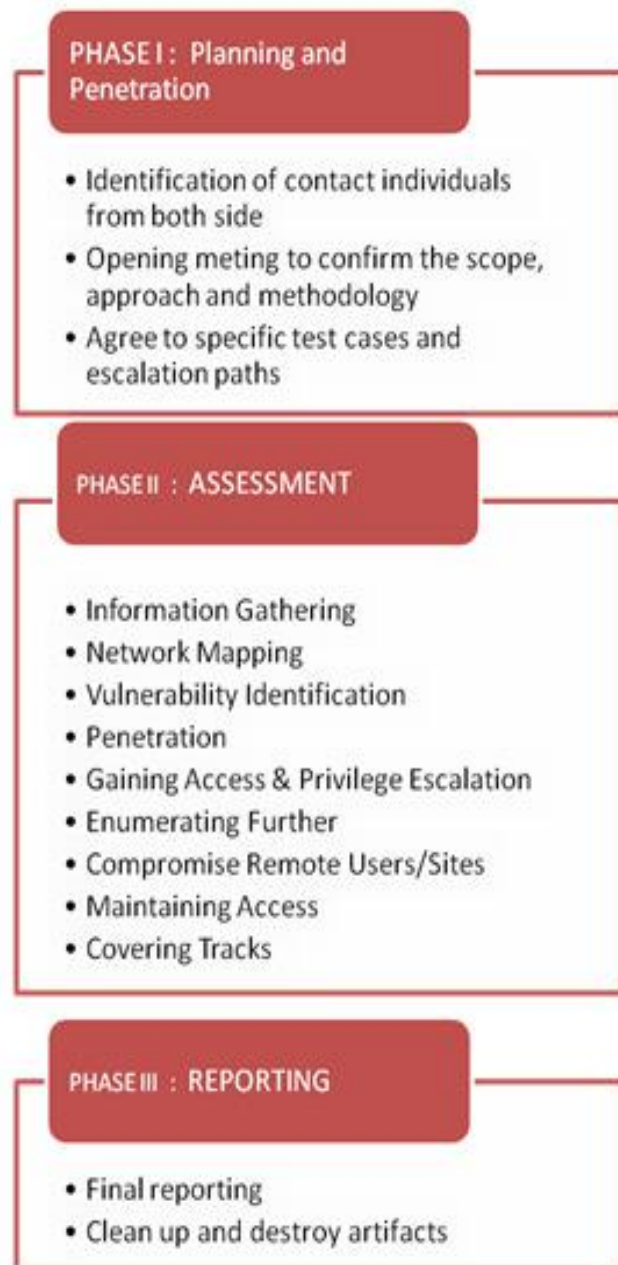


Figure 2: The Phases of Penetration Testing (ISSAF) Standard

The final standard to be considered is the PTES, the Penetration Testing Execution Standard is a completely new standard that started to be developed in 2010. One of its main features is that the industry experts in specific areas have developed it [9]. the steps covered in the PTES are shown in (Figure.3) [6].

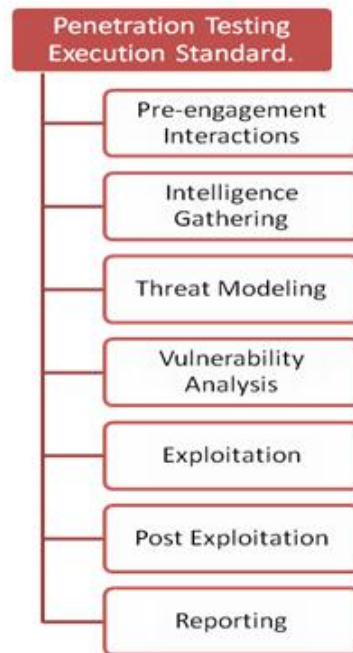


Figure 3: Steps of Penetration Testing Execution Standard

### 3. MANUAL VS. AUTOMATED PENETRATION TESTING

Until recently, Penetration testing has been restricted to advanced security specialist that having many years of relevant experience to do the complex manual process. but in fact, the proficient penetration testers are not highly available and the manual process is time and money consuming. A team of experts can gather to build a professional automated tool that can be a combination of experience of the expert penetration testers. so that the non-expert users can substitute the penetration team with the automated tools to get an inclusive view of the security situation on the organization's system.

The table below, summaries the comparison between manual and automated penetration testing:

Table 1: Comparison of manual and automated testing[1] [10]

	Automated	Manual
<b>Testing process</b>	Fast, standard process; Easily repeatable tests;	Manual, non-standard process; capital intensive; High cost of customization;
<b>Vulnerability /attack Database management</b>	Attack database is maintained and updated attack codes are written for a variety of platforms;	Maintenance of database is manual; Need o rely on public database; Need re-write attack code for functioning across different platforms;

<b>Exploit Development and Management</b>	Product vendor develops and maintains all exploits. Exploits are continually updated for maximum effectiveness. Exploits are professionally developed, thoroughly tested, and safe to run. Exploits are written and optimized for a variety of platforms and attack vectors	Developing and maintaining an exploit database is time-consuming and requires significant expertise. Public exploits are suspect and can be unsafe to run. Re-writing and porting code is necessary for cross platform functionality.
<b>Reporting</b>	Reports are automated and customized	Requires collecting the data manually
<b>Cleanup</b>	Automated testing products offer clean-u solutions	The tester has to manually undo the changes to the system every time vulnerabilities found
<b>Network modification</b>	System remain unchanged.	Often results in numerous system modification
<b>Logging/ Auditing</b>	Automatically records a detailed record of all activity.	Slow, cumbersome, often inaccurate process
<b>Training</b>	Training for automated tools is easier than manual testing	Testers need to learn non-standard ways of testing ; training can be customized and is time consuming

#### 4. CURRENT AUTOMATED PENETRATON TESTING OVERVIEW

Recently, penetration testing has been used to find the vulnerabilities exists in the system to know how to mitigate them. the test usually simulates various types of attacks on the target system. by this test, the administrator will have an organized and controlled way to identify the security shortcomings. The resources and time needed for comprehensive testing will make penetration testing price intensive. Consequently, such tests are sometimes solely performed throughout necessary milestones. during [5] project have been automated the penetration testing method for many protocol-based attacks. their automated penetration testing, application covers many attacks which support hypertext transfer protocol (HTTP), SIP and TCP/IP. the target of this work is to supply a quick, reliable and automated testing tool, that is additionally easier to use than existing tools.

In research [6], The purpose behind this research was to contribute a tool to the community that may be won't to improve the potency of current penetration testing companies, so that they will expand coverage of the testing to grant customers a additional in depth read of their current security ways and wherever they have to enhance. the most purpose behind this tool is to prove that automated testing isn't a hindrance to the security community however is very a tool that should be leveraged throughout testing.



In the other hand project [11] was developed to facilitate the vulnerability analysis and penetration testing in Indian banks. It's a big threat to the bank to do the penetration testing an vulnerability assessment using third party tools. this approach is fully automated and interactive so that it doesn't require a high experience and technical skills from the users. the tool has been developed using python libraries and without any third party software's, it's a reliable option to find the vulnerabilities associated to the applications and services running on the target system. After that, the tool produces a vulnerability list with the severity level associated to each vulnerability. it also used to detect the SQLI vulnerability on the target system.

In [3], the known mathematical model of partially observable Markov decision processes (POMDP) have been used to tackle the problem of the research. the solution proposed automatically produce generation of multi-step plans for a penetration test, the plans are robust to uncertainty during execution. this work focuses on remote test with uncertainty in both information gathering and exploit actions, by developing probabilistic metrics to find the effective probability that an exploit can be executed and the overall probability that the attacker can successfully execute it a summary for the comparison between these approaches can be found in table 2. Paper [5] will be considered as A , paper [6] will be considered as B while paper [11] will be C and paper [3] will be D.

Table 2: Comparison between current methodologies of automated penetration testing

	Target	Tools	Phases	Method of implementation	Aim
A	HTTP / TCP/IP and session initiation protocol (SIP) attacks	Hping3 to perform TCP DOS attack	Input parameters based on the web interface then routing the params to appropriate module then finally implement the attack	The application was developed using PHP and the attack scripts is implemented using JAVA and shell scripting , the design implemented on Linux	Perform automated easy to use with web interface, penetration testing toolkit
B	All protocols and services	Harvester, Metagoofil, NMAP, ZAP, Metasploit, Nessus (pynessus)	Insert arguments , run scanning tools then parse the output from these tools and finally exploit with Metasploit and start the manual process if needed	Uses script to link the tools to each other's and to parse the output from them.	Optimizing the process by automating the running of any tool that is used commonly on the penetration test
C	Database	No third party tools has been used here	Information gathering, scanning, then	Developed using core python packages/Libraries and	Find user credentials , Email ID and

			vulnerability detection and mapping and in the final step Exploitation and report generation is done	no third party software has been used	other details from the database using SQLI
D	All services and protocols	Exploiting tools	Information gathering then vulnerability assessment and the final step is penetration test planning	Used partially observable marker decision process ( POMDPs) and demonstrate the use of an effective approximation algorithm that satisfies the performance requirement of penetration testing planning , then a script is used to link the components	Find a way to do remote penetration testing with uncertainty of tools used.

## 5. CONCLUSION AND FUTURE WORK

Many organizations need penetration testing to discover the most vulnerabilities that have in their system. To apply the penetration test, there are two approaches that the organizations used to discover the bugs, one is automated penetration test and the other is manual penetration test. The automated pen test is the easiest way to figure out the whole vulnerabilities in the system by implementing a tool that has some patterns to find the vulnerabilities. While the manual test is the way to discover the vulnerabilities manually through analyzing the system and distinguish the abnormal behavior.

Hence, this paper has been done to shows the importance of the penetration testing as well as the importance of automating this process. Additionally, some standards in the penetration testing have been highlighted to help the researchers find the suitable standards to use. Even more, the comparison between the manual and automated penetration testing has been provided in term of the testing process, vulnerability and attack database management, exploit development and management reporting, clean up, network modification, logging, and training. And the result shows that the automated penetration testing is better than the manual penetration in all of the above process, except finding the new or zero day exploits. So that many organizations may go for the automated approach because it seems the better and cheaper way to maintain security in the systems as most of the vulnerabilities that the attackers used to exploit the system are well defined in the automated tools.

Although writing own exploits may be time-consuming as well as ineffective in terms of money. But, the attackers can conceal their activity through their own scripts. Thus, the automated tools still have limitations and vulnerabilities.

To the best of our knowledge, this research is a step forward to the other researchers who interested in the automated penetration testing. The next step is to study the impact of the penetration testing toward the hunting threats. Even more, to study the applicability of building automated tool that takes into consideration the general limitations in the current automated tools.

## REFERENCES

- [1] Y. Stefinko, A. Piskozub and R. Banakh, "Manual and automated penetration testing. Benefits and drawbacks. Modern tendency," 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), Lviv, 2016, pp. 488-491. doi: 10.1109/TCSET.2016.7452095
- [2] Xue Qiu, Shuguang Wang, Qiong Jia, Chunhe Xia and Qingxin Xia, "An automated method of penetration testing," 2014 IEEE Computers, Communications and IT Applications Conference, Beijing, 2014, pp. 211-216. doi: 10.1109/ComComAp.2014.7017198
- [3] L. Greenwald and R. Shanley, "Automated planning for remote penetration testing," MILCOM 2009 - 2009 IEEE Military Communications Conference, Boston, MA, 2009, pp. 1-7. doi: 10.1109/MILCOM.2009.5379852
- [4] Gula, Ron. "Broadening the Scope of Penetration Testing Techniques." Jul. 1999. URL: [www.forum-intrusion.com/archive/ENTRASYS.pdf](http://www.forum-intrusion.com/archive/ENTRASYS.pdf) (6/14/12)
- [5] Samant, Neha. Automated penetration testing. Diss. San Jose State University, 2011.
- [6] K. P. Haubris and J. J. Pauli, "Improving the Efficiency and Effectiveness of Penetration Test Automation," 2013 10th International Conference on Information Technology: New Generations, Las Vegas, NV, 2013, pp. 387-391. doi: 10.1109/ITNG.2013.135
- [7] Souppaya, Karen Scarfone Murugiah, Amanda Cody, and Angela Orebaugh. "Technical Guide to Information Security Testing and Assessment." Recommendations of the National Institute of Standards and Technology (2008).
- [8] "Open Information Systems Security Group", Information systems security assessment framework, 2006.
- [9] Pentest-standard.org. (2018). The Penetration Testing Execution Standard. [online] Available at: [http://www.pentest-standard.org/index.php/Main\\_Page](http://www.pentest-standard.org/index.php/Main_Page) [Accessed 31 Mar. 2018].
- [10] Mirjalili, Mahin, Alireza Nowroozi, and Mitra Alidoosti. "A survey on web penetration test." International Journal in Advances in Computer Science 3.6 (2014).
- [11] Shah, Sugandh, and B. M. Mehtre. "An automated approach to Vulnerability Assessment and Penetration Testing using Net-Nirikshak 1.0." Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference on. IEEE, 2014.

*INTENTIONAL BLANK*

# EXACT SOLUTIONS OF A FAMILY OF HIGHER-DIMENSIONAL SPACE-TIME FRACTIONAL KDV-TYPE EQUATIONS

MOHAMMED O. AL-AMR

Department of Mathematics, College of Computer Sciences and Mathematics,  
University of Mosul, Mosul 41002, Iraq

## ABSTRACT

*In this paper, based on the definition of conformable fractional derivative, the functional variable method (FVM) is proposed to seek the exact traveling wave solutions of two higher-dimensional space-time fractional KdV-type equations in mathematical physics, namely the (3+1)-dimensional space-time fractional Zakharov-Kuznetsov (ZK) equation and the (2+1)-dimensional space-time fractional Generalized Zakharov-Kuznetsov-Benjamin-Bona-Mahony (GZK-BBM) equation. Some new solutions are procured and depicted. These solutions, which contain kink-shaped, singular kink, bell-shaped soliton, singular soliton and periodic wave solutions, have many potential applications in mathematical physics and engineering. The simplicity and reliability of the proposed method is verified.*

## KEYWORDS

*Functional Variable Method, Fractional Partial Differential Equations, Exact Solutions, Conformable Fractional Derivative*

## 1. INTRODUCTION

In recent years, Fractional partial differential equations (FPDEs) have been extensively utilized to model complex physical phenomena that arise in various aspects of science and engineering, such as applied mathematics, physics, chemistry, biology, signal processing, control theory, finance and fractional dynamics [1,2]. The analytical solutions of FPDEs play a significant role in the study of nonlinear physical phenomena. Therefore, the efficient approaches to construct the solutions of FPDEs have attracted great interest by several groups of researchers. A large collection of analytical and computational methods has been introduced for this reason, for example the exp-function method [3,4], Adomian decomposition method [5], the  $(G'/G)$ -expansion method [6], the first integral method [7,8], the variational iteration method [9], the sub-equation method [10,11], the modified simple equation method [12], Jacobi elliptic function expansion method [13], the generalized Kudryashov method [14,15] and so on. One of the most powerful methods for seeking analytical solutions of nonlinear differential equations is the functional variable method, which was first proposed by Zerarka et al. [16,17] in 2010. It has received much interest since it has been employed to solve a wide class of problems by many authors [18-22]. The main advantage of this method over other existing methods is its capability

to reduce the size of computations during the solution procedure. Therefore, it can be applied without using any symbolic computation software.

As one of the most well-known nonlinear dispersive equations, the Korteweg-de Vries (KdV) equation has attracted much attention by many researchers in the scientific community due to its significant role in various scientific disciplines. It describes a variety of important nonlinear phenomena, including finite amplitude dispersive wave phenomena, acoustic waves in a harmonic crystal and ion-acoustic waves in plasmas [23]. Several variations of this equation have been introduced in the literature. The (3+1)-dimensional Zakharov-Kuznetsov (ZK) equation was derived as a three-dimensional generalization of the KdV equation, which arises as a model for the propagation of nonlinear plasma-acoustic waves in the isothermal multi-component magnetized plasma [24,25]. If the nonlinear dispersion in KdV equation is incorporated, the Benjamin-Bona-Mahony (BBM) equation arises to describe a propagation of long waves. The (2+1)-dimensional Generalized Zakharov-Kuznetsov-Benjamin-Bona-Mahony (GZK-BBM) equation was developed by Wazwaz [26] as a combination of the well-known Benjamin-Bona-Mahony (BBM) equation with the ZK equation. It arises as a description of gravity water waves in the long-wave regime. Therefore, it is very interesting to examine the traveling wave solutions of KdV-type equations. It is worthwhile to mention that the (3+1)-dimensional space-time fractional ZK equation and the (2+1)-dimensional space-time fractional GZK-BBM equation have not been solved yet by using any existing analytical method.

There are many different definitions for fractional differential equations in fractional calculus; among these definitions are Riemann–Liouville, Grünwald–Letnikov, Caputo, Weyl, Marchaud, Hadamard, Canavati, Davidson–Essex, Riesz–Fischer, Jumarie fractional derivatives and so on [2,27]. However, these definitions have some shortcomings. For instance, they do not satisfy the product rule, the quotient rule and the chain rule for derivative operations. To overcome these drawbacks, Khalil et al. [28] introduced a completely new definition of the fractional derivative, which is more natural and fruitful than previous ones, called conformable fractional derivative.

The present paper is devoted to suggest the functional variable method for constructing new exact solutions of two higher-dimensional space-time fractional KdV-related equations, namely the (3+1)-dimensional space-time fractional ZK equation and the (2+1)-dimensional space-time fractional GZK-BBM equation. The fractional derivatives are presented in terms of the conformable sense. To the best of our knowledge, these equations have not been investigated previously by using the functional variable method in the sense of conformable derivative.

The rest of the paper is organized as follows: In Section 2, we describe some relevant materials and methods. In Section 3, the proposed approach is applied to establish the exact solutions of the underlying equations. The graphical representations of the obtained solutions are provided in Section 4, Results and discussion are presented in Section 5. Finally, conclusions are given in Section 6.

## 2. MATERIALS AND METHODS

### 2.1. Conformable fractional derivative and its properties

In this subsection, we present some basic definitions and properties of the conformable fractional calculus. Suppose a function  $f : [0, \infty) \rightarrow \mathbb{R}$ , then, the conformable fractional derivative of order  $\alpha$  is defined as follows [28,29]:

$$T_{\alpha}(f)(t) = \lim_{\varepsilon \rightarrow 0} \frac{f(t + \varepsilon t^{1-\alpha}) - f(t)}{\varepsilon}, \quad (1)$$

in which  $t > 0$  and  $0 < \alpha \leq 1$ . If  $f$  is  $\alpha$ -differentiable in some  $(0, a)$ ,  $a > 0$ , and  $\lim_{t \rightarrow 0^+} f^{(\alpha)}(t)$  exists, then  $f^{(\alpha)}(0) = \lim_{t \rightarrow 0^+} f^{(\alpha)}(t)$ .

Now, we summarize some useful properties of the conformable derivative as follows [28-30]:

- (i)  $T_{\alpha}(af + bg) = aT_{\alpha}(f) + bT_{\alpha}(g)$ , for all  $a, b \in \mathbb{R}$ .
- (ii)  $T_{\alpha}(t^p) = pt^{p-\alpha}$ , for all  $p \in \mathbb{R}$ .
- (iii)  $T_{\alpha}(fg) = fT_{\alpha}(g) + gT_{\alpha}(f)$ .
- (iv)  $T_{\alpha}\left(\frac{f}{g}\right) = \frac{gT_{\alpha}(f) - fT_{\alpha}(g)}{g^2}$ .
- (v)  $T_{\alpha}(\lambda) = 0$ , where  $\lambda$  is a constant.
- (vi) If  $f$  is differentiable, then  $T_{\alpha}(f)(t) = t^{1-\alpha} \frac{df}{dt}$ .
- (vii) If  $f, g$  are differential functions, then  $T_{\alpha}(f \circ g)(t) = t^{1-\alpha} g'(t) f'(g(t))$ .

Moreover, some conformable fractional derivatives of certain functions can be found in [28]. The abovementioned properties will be utilized further in the forthcoming sections.

## 2.2. Description of the functional variable method

Consider the following general FPDE with four independent variables:

$$P(u, \frac{\partial^{\alpha} u}{\partial t^{\alpha}}, \frac{\partial^{\alpha} u}{\partial x^{\alpha}}, \frac{\partial^{\alpha} u}{\partial y^{\alpha}}, \frac{\partial^{\alpha} u}{\partial z^{\alpha}}, \frac{\partial^{2\alpha} u}{\partial t^{2\alpha}}, \frac{\partial^{2\alpha} u}{\partial x^{2\alpha}}, \dots) = 0, \quad 0 < \alpha \leq 1 \quad (2)$$

where  $P$  is a polynomial of  $u(x, y, z, t)$  and its fractional partial derivatives, in which the highest order derivatives and the nonlinear terms are involved.

The foremost steps of the FVM can be outlined as follows [18,19]:

Step 1: To find the exact solution of Eq. (2), we use the fractional complex transformation

$$u(x, y, z, t) = u(\xi), \quad \xi = \frac{k_1 x^{\alpha}}{\alpha} + \frac{k_2 y^{\alpha}}{\alpha} + \frac{k_3 z^{\alpha}}{\alpha} + \frac{ct^{\alpha}}{\alpha}, \quad (3)$$

where  $k_1, k_2, k_3$  and  $c$  are nonzero arbitrary constants, to convert Eq. (2) into the following ordinary differential equation (ODE) of integer order:

$$\tilde{P}(u, cu', k_1 u', k_2 u', k_3 u', c^2 u'', k_1^2 u'', \dots) = 0, \quad (4)$$

where  $\tilde{P}$  is a polynomial in  $u(\xi)$  and its total derivatives with respect to  $\xi$ .

Step 2: Let us make a transformation in which the unknown function  $u(\xi)$  is considered as a functional variable in the form

$$u_\xi = F(u), \quad (5)$$

It is easy to find some higher order derivatives of  $u(\xi)$  as follows:

$$\begin{aligned} u_{\xi\xi} &= FF' = \frac{1}{2} (F^2)', \\ u_{\xi\xi\xi} &= \frac{1}{2} (F^2)'' F = \frac{1}{2} (F^2)'' \sqrt{F^2}, \\ u_{\xi\xi\xi\xi} &= \frac{1}{2} \left( (F^2)''' F^2 + \frac{1}{2} (F^2)'' (F^2)' \right), \end{aligned} \quad (6)$$

and so on, where the prime denotes the derivative with respect to  $u$ .

Step 3: We substitute Eqs. (5) and (6) into Eq. (4) to reduce it to the following ODE:

$$R(u, F, F', F'', \dots) = 0. \quad (7)$$

Step 4: After integration, Eq. (7) provides the expression of  $F$ , and this in turn together with Eq. (5) gives the appropriate solutions to the original equation.

### 3. APPLICATIONS

In this section, we apply the functional variable method, which described in the previous section, to look for the exact solutions of two higher-dimensional space-time fractional equations of KdV-type.

#### 3.1 The (3+1)-dimensional space-time fractional ZK equation

Consider the (3+1)-dimensional space-time fractional ZK equation [24,25]

$$D_t^\alpha u + au D_x^\alpha u + D_x^{2\alpha} u + D_y^{2\alpha} u + D_z^{2\alpha} u = 0, \quad (8)$$

where  $0 < \alpha \leq 1$  and  $a$  is a nonzero constant.

To investigate Eq. (8) using the FVM, we use the fractional complex transformation given by Eq. (3) to reduce Eq. (8) into the following ODE:

$$cu_\xi + ak_1 uu_\xi + (k_1^2 + k_2^2 + k_3^2) u_{\xi\xi} = 0, \quad (9)$$

Integrating once w.r.t.  $\xi$  and setting the constant of integration to zero, yields

$$cu + \frac{ak_1}{2} u^2 + (k_1^2 + k_2^2 + k_3^2) u_\xi = 0, \quad (10)$$



Substituting Eq. (5) into Eq. (10), the function  $F(u)$  reads

$$F(u) = -\frac{c}{k_1^2 + k_2^2 + k_3^2} u \left( 1 + \frac{ak_1}{2c} u \right), \quad (11)$$

Separating the variables in Eq. (11) and then integrating, we obtain

$$\int \frac{-du}{u \left( 1 + \frac{ak_1}{2c} u \right)} = \frac{c}{k_1^2 + k_2^2 + k_3^2} (\xi + \xi_0), \quad (12)$$

where  $\xi_0$  is a constant of integration. After completing the integration of Eq. (12), we get the following exact solutions:

$$u_1(\xi) = -\frac{c}{ak_1} \left( 1 - \tanh \left( \frac{c}{2(k_1^2 + k_2^2 + k_3^2)} (\xi + \xi_0) \right) \right), \quad (13)$$

$$u_2(\xi) = -\frac{c}{ak_1} \left( 1 - \coth \left( \frac{c}{2(k_1^2 + k_2^2 + k_3^2)} (\xi + \xi_0) \right) \right). \quad (14)$$

$$\text{where } \xi = \frac{k_1 x^\alpha}{\alpha} + \frac{k_2 y^\alpha}{\alpha} + \frac{k_3 z^\alpha}{\alpha} + \frac{ct^\alpha}{\alpha}.$$

### 3.2 The (2+1)-dimensional space–time fractional GZK–BBM equation

Consider the (2+1)-dimensional space–time fractional GZK-BBM equation in the form [26]

$$D_t^\alpha u + D_x^\alpha u + aD_x^\alpha u^n + bD_x^\alpha (D_x^\alpha D_t^\alpha u + D_y^{2\alpha} u) = 0, \quad n > 1 \quad (15)$$

where  $0 < \alpha \leq 1$  and  $a, b$  are nonzero constants.

To apply the FVM for Eq. (15). We exploit the fractional complex transformation

$$u(x, y, t) = u(\xi), \quad \xi = \frac{k_1 x^\alpha}{\alpha} + \frac{k_2 y^\alpha}{\alpha} + \frac{ct^\alpha}{\alpha}, \quad (16)$$

to convert Eq. (15) into the following ODE:

$$(c + k_1)u_\xi + ak_1(u^n)_\xi + bk_1(k_1 cu_{\xi\xi} + k_2^2 u_{\xi\xi})_\xi = 0, \quad (17)$$

Integrating once w.r.t.  $\xi$  with zero constant of integration, we obtain

$$(c + k_1)u + ak_1 u^n + bk_1(k_1 c + k_2^2)u_{\xi\xi} = 0, \quad (18)$$

Substituting Eqs. (6) into Eq. (18), yields

$$(F^2)' = -\frac{2}{bk_1(k_1c + k_2^2)} \left[ (c + k_1)u + ak_1u^n \right], \quad (19)$$

Integrating Eq. (19) w.r.t.  $u$ , we deduce the expression of the function  $F(u)$  as follows

$$F(u) = \sqrt{\frac{-(c + k_1)}{bk_1(k_1c + k_2^2)}} u \sqrt{1 + \frac{2ak_1}{(c + k_1)(n+1)} u^{n-1}}, \quad (20)$$

Separating the variables in Eq. (20) and then integrating, we obtain

$$\int \frac{du}{u \sqrt{1 + \frac{2ak_1}{(c + k_1)(n+1)} u^{n-1}}} = \sqrt{\frac{-(c + k_1)}{bk_1(k_1c + k_2^2)}} (\xi + \xi_0), \quad (21)$$

where  $\xi_0$  is a constant of integration. After completing the integration of Eq. (21), we can simply attain the following exact solutions:

(i) If  $\frac{c + k_1}{bk_1(k_1c + k_2^2)} < 0$ , we have the following hyperbolic solutions:

$$u_1(\xi) = \left\{ -\frac{(c + k_1)(n+1)}{2ak_1} \operatorname{sech}^2 \left( \frac{n-1}{2} \sqrt{\frac{-(c + k_1)}{bk_1(k_1c + k_2^2)}} (\xi + \xi_0) \right) \right\}^{\frac{1}{n-1}}, \quad (22)$$

$$u_2(\xi) = \left\{ \frac{(c + k_1)(n+1)}{2ak_1} \operatorname{csch}^2 \left( \frac{n-1}{2} \sqrt{\frac{-(c + k_1)}{bk_1(k_1c + k_2^2)}} (\xi + \xi_0) \right) \right\}^{\frac{1}{n-1}}, \quad (23)$$

where  $\xi = \frac{k_1x^\alpha}{\alpha} + \frac{k_2y^\alpha}{\alpha} + \frac{ct^\alpha}{\alpha}$ .

(ii) If  $\frac{c + k_1}{bk_1(k_1c + k_2^2)} > 0$ , we have the following trigonometric solutions:

$$u_3(\xi) = \left\{ -\frac{(c + k_1)(n+1)}{2ak_1} \sec^2 \left( \frac{n-1}{2} \sqrt{\frac{(c + k_1)}{bk_1(k_1c + k_2^2)}} (\xi + \xi_0) \right) \right\}^{\frac{1}{n-1}}, \quad (24)$$

$$u_4(\xi) = \left\{ -\frac{(c + k_1)(n+1)}{2ak_1} \csc^2 \left( \frac{n-1}{2} \sqrt{\frac{(c + k_1)}{bk_1(k_1c + k_2^2)}} (\xi + \xi_0) \right) \right\}^{\frac{1}{n-1}}. \quad (25)$$

where  $\xi = \frac{k_1 x^\alpha}{\alpha} + \frac{k_2 y^\alpha}{\alpha} + \frac{ct^\alpha}{\alpha}$ .

#### 4. GRAPHICAL ILLUSTRATIONS

In this section, with the aid of Maple software, we show the graphical representation of some results in Figs. 1-3 by assigning appropriate values to the unknown parameters in order to visualize the mechanism of Eqs. (8) and (15). Some physical interpretations are also presented.

##### 4.1 The (3+1)-dimensional space–time fractional ZK equation

The profiles of the kink-shaped solution  $u_1(\xi)$  given by Eq. (13) is shown in Fig. 1 when  $y = 0, z = 1, a = 1, k_1 = 1.5, k_2 = 0.25, k_3 = 1, c = -2, \xi_0 = 0$  for various values of  $\alpha$ . We can observe that when the fractional derivative order  $\alpha$  increased, the shape is closer to the known kink wave as the velocity of the propagation wave decreases. The kink wave keeps its height for various values of  $\alpha$ . It should also be pointed out that the solution  $u_2(\xi)$  given by Eq. (14) is a singular kink solution.

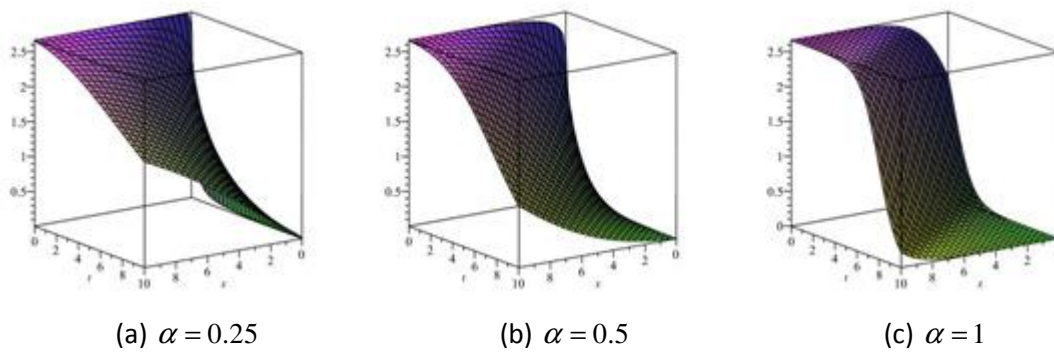


Figure 1. The kink solution corresponding to Eq. (13) for various values of  $\alpha$

##### 4.2 The (2+1)-dimensional space–time fractional GZK-BBM equation

The dynamics of the singular soliton solution  $u_2(\xi)$  given by Eq. (23) is shown in Fig. 2 when  $y = 1, a = 1, b = -2, k_1 = 1.25, k_2 = -4, c = 2, n = 4, \xi_0 = 0$  for various values of  $\alpha$ . When  $\alpha$  increased, the height of the wave changes as the velocity of the wave propagation decreases. Fig. 3 shows the motions of the periodic wave solution  $u_3(\xi)$  given by Eq. (24) when  $y = 0, a = 1, b = 0.5, k_1 = 0.5, k_2 = 0.25, c = -2, n = 4, \xi_0 = 0$  for various values of  $\alpha$ . When  $\alpha$  increased, the height of the wave becomes lower as the velocity of the wave propagation decreases. It should also be mentioned that the solution  $u_1(\xi)$  given by Eq. (22) is a bell-shaped soliton solution.

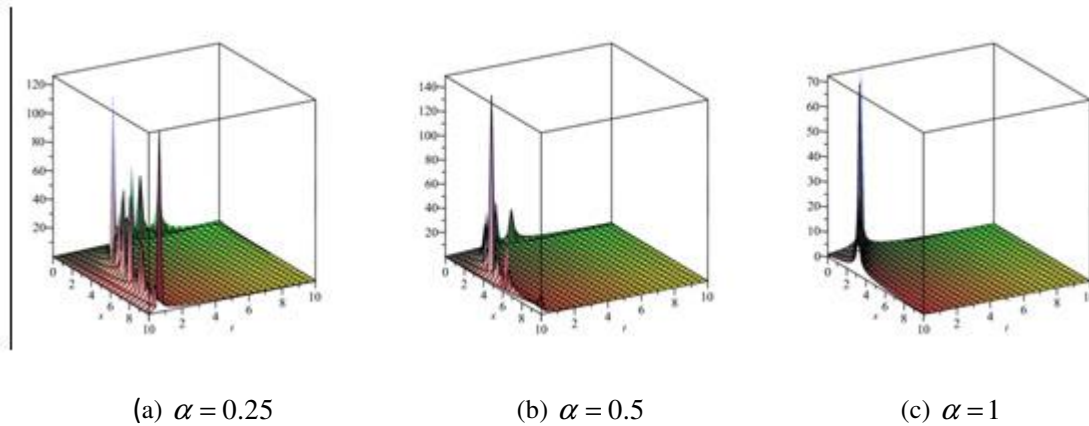


Figure 2. The singular soliton solution corresponding to Eq. (23) for various values of  $\alpha$

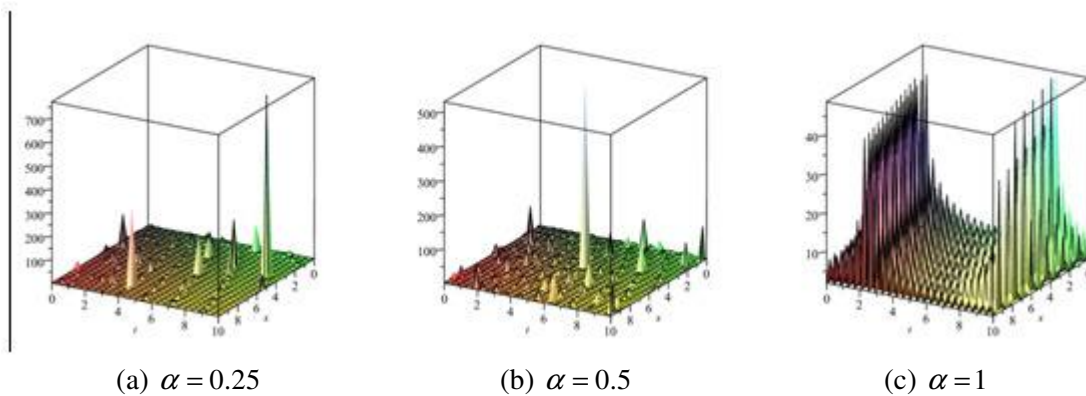


Figure 3. The periodic solution corresponding to Eq. (24) for various values of  $\alpha$

## 5. RESULTS AND DISCUSSION

For the first time, the analytical solutions of the (3+1)-dimensional space–time fractional ZK equation and the (2+1)-dimensional space–time fractional GZK-BBM equation have been attained via the functional variable method, the fractional derivative has been described in the conformable sense. Consequently, we deduce that our solutions (13), (14), (22)-(25) are new and not discussed heretofore. It is remarkable that the obtained solutions in this article have potential physical meaning for the underlying equations. In addition to the physical meaning, these solutions can be used to measure the accuracy of numerical results and to help in the study of stability analysis.

## 6. CONCLUSIONS

In this paper, we have successfully executed the functional variable method to attain new exact traveling wave solutions of a family of higher-dimensional space-time fractional KdV-type

equations arising in mathematical physics, namely the (3+1)-dimensional space–time fractional ZK equation and the (2+1)-dimensional space–time fractional GZK-BBM equation. Two types of solutions including hyperbolic function solutions and trigonometric function solutions are obtained and numerically simulated in Figs. 1-3. The obtained solutions are significant to reveal the inner mechanism of the nonlinear physical phenomena that described by the aforementioned equations. It is shown that the FVM is straightforward, powerful and can be extended to handle many other higher-dimensional fractional partial differential equations as it maintains the reduced volume of computational work. With the aid of the Maple, we have verified our results.

## REFERENCES

- [1] Samko, S.G., Kilbas, A.A., Marichev, O.I., (1993) Fractional Integrals and Derivatives Theory and Applications. Gordon and Breach, New York.
- [2] Podlubny, I., (1999) Fractional Differential Equations. Academic Press, San Diego.
- [3] Zhang, S., Zong, Q.A., Liu, D., Gao, Q., (2010) A generalized exp-function method for fractional Riccati differential equations. Commun. Fract. Calc. 1, 48-51.
- [4] Bekir, A., Guner, O., Cevikel, A.C., (2013) Fractional complex transform and exp-function methods for fractional differential equations. Abstr. Appl. Anal. 2013, 426462.
- [5] El-Sayed, A.M.A., Rida, S.Z., Arafa, A.A.M., (2009) Exact solutions of fractional-order biological population model. Commun. Theor. Phys. 52, 992 -996.
- [6] Shang, N., Zheng, B., (2013) Exact solutions for three fractional partial differential equations by the method. Int. J. Appl. Math. 43 (3), 114-119.
- [7] Ekici, M., Mirzazadeh, M., Eslami, M., Zhou, Q., Moshokoa, S.P., Biswas, A., Belic, M., (2016) Optical soliton perturbation with fractional-temporal evolution by first integral method with conformable fractional derivatives. Optik 127, 10659–10669.
- [8] Eslami, M., Rezazadeh, H., (2016) The first integral method for Wu–Zhang system with conformable time-fractional derivative. Calcolo 53, 475–485.
- [9] Inc, M., (2008) The approximate and exact solutions of the space- and time-fractional Burgers equations with initial conditions by variational iteration method. J. Math. Anal. Appl. 345, 476-484.
- [10] Aminikhah, H., Sheikhan, A.H.R., Rezazadeh, H., (2016) Sub-equation method for the fractional regularized long-wave equations with conformable fractional derivatives. Scientia Iranica B 23 (3), 1048-1054.
- [11] Zheng, B., Wen, C., (2013) Exact solutions for fractional partial differential equations by a new fractional sub-equation method. Adv. Difference Equ. 2013, 199.
- [12] Kaplan, M., Bekir, A., Akbulut, A., Aksoy, E., (2015) The modified simple equation method for nonlinear fractional differential equations. Rom. Jour. Phys. 60, 1374-1383.
- [13] Tasbozan, O., Çenesiz, Y., Kurt, A., (2016) New solutions for conformable fractional Boussinesq and combined KdV-mKdV equations using Jacobi elliptic function expansion method. Eur. Phys. J. Plus 131, 244.

- [14] Demiray, S.T., Pandir, Y., Bulut, H., (2014) The investigation of exact solutions of nonlinear time-fractional Klein-Gordon equation by using generalized Kudryashov method. AIP Conf. Proc. 1637, 283 -289.
- [15] Demiray, S.T., Pandir, Y., Bulut, H., (2014) Generalized Kudryashov method for time-fractional differential equations. Abstr. Appl. Anal. 2014, 901540.
- [16] Zerarka, A., Ouamane, S., Attaf, A., (2010) On the functional variable method for finding exact solutions to a class of wave equations. App. Math. and Com. 217, 2897–2904.
- [17] Zerarka, A., Ouamane, S., (2010) Application of the functional variable method to a class of nonlinear wave equations. World J. Model. Simul. 6 (2), 150-160.
- [18] Liu, W., Chen, K., (2013) The functional variable method for finding exact solutions of some nonlinear time-fractional differential equations. Pramana J. Phys., 81 (3), 377–384.
- [19] Bekir, A., Güner, Ö., Aksoy, E., Pandir, Y., (2015) Functional variable method for the nonlinear fractional differential equations. AIP Conf. Proc. 1648, 730001.
- [20] Khan, K., Akbar, M.A., (2015) Study of functional variable method for finding exact solutions of nonlinear evolution equations. Walailak J. Sci & Tech. 12 (11), 1031-1042.
- [21] Mirzazadeh, M., Eslami, M., (2013) Exact solutions for nonlinear variants of Kadomtsev-Petviashvili (n,n) equation using functional variable method. Pramana J. Phys. 81, 911-24.
- [22] Bekir, A., San, S., (2013) Periodic, hyperbolic and rational function solutions of nonlinear wave equations. Appl. Math. Inf. Sci. Lett. 1 (3), 97-101.
- [23] Wazwaz, A.M., (2009) Partial Differential Equations and Solitary Waves Theory (Higher Education Press, Beijing) P. 503.
- [24] Zakharov, V.E., Kuznetsov, E.A., (1974) Three-dimensional solitons. Sov. Phys. 39, 285-286.
- [25] Das, G.C., Sarma, J., Gao, Y.T., Uberoi, C., (2000) Dynamical behavior of the soliton formation and propagation in magnetized plasma. Phys. Plasmas 7, 2374- 2380.
- [26] Wazwaz, A.M., (2005) Compact and noncompact physical structures for the ZK–BBM equation. App. Math. Comput. 169, 713–725.
- [27] Oliveira, E.C. & Machado, J.A.T., (2014) A review of definitions of fractional derivatives and Integral, Math. Probl. Eng. 2014 238459.
- [28] Khalil, R., Horani, AL.M., Yousef, A., Sababheh, M., (2014) A new definition of fractional derivative. J. Comput. Appl. Math. 264, 65-70.
- [29] Abu Hammad, M., Khalil, R., (2014) Conformable fractional heat differential equation. Int. J. Pure Appl. Math. 94 (2), 215-221.
- [30] Abdeljawad, T., (2015) On conformable fractional calculus. J. Comput. Appl. Math. 279 (1), 57-66.

**AUTHOR**

**Mohammed O. Al-Amr** was born on January 29th, 1986 in Mosul, Iraq. He received his B.Sc. in Mathematics from University of Mosul in 2007. He received his M.Sc. in Mathematics from University of Mosul in 2013 and studied in the field of “Numerical Analysis”. Since 2013, he has been an assistant lecturer at University of Mosul. He published many papers in reputable scientific journals. He serves as a reviewer and editorial member of many scientific journals. He is a member of many international scientific associations. His main research interests are numerical analysis, partial differential equations, semi-analytical methods, stability analysis, traveling wave analysis, theory of solitons.



*INTENTIONAL BLANK*



# IMPORTANCE OF VERB SUFFIX MAPPING IN DISCOURSE TRANSLATION SYSTEM

Suryakanthi Tangirala

Faculty of Business, University of Botswana, Gaborone, Botswana

## ABSTRACT

*This paper discusses the importance of verb suffix mapping in Discourse translation system. In discourse translation, the crucial step is Anaphora resolution and generation. In Anaphora resolution, cohesion links like pronouns are identified between portions of text. These binders make the text cohesive by referring to nouns appearing in the previous sentences or nouns appearing in sentences after them. In Machine Translation systems, to convert the source language sentences into meaningful target language sentences the verb suffixes should be changed as per the cohesion links identified. This step of translation process is emphasized in the present paper. Specifically, the discussion is on how the verbs change according to the subjects and anaphors. To explain the concept, English is used as the source language (SL) and an Indian language Telugu is used as Target language (TL).*

## KEYWORDS

*MT: Machine Translation, SL-Source language, TL: target Language, POS-Parts of Speech, GNP-Gender Number Person.*

## 1. INTRODUCTION

Language is the medium of communication. Language apart from being a communication medium is a powerful source of information exchange. Different people use different languages. Though English is the globally accepted language, mostly people understand the things better in their native language. In this era of globalization, information sharing is very important. If machine translators are made, languages will not be a barrier any more. Web is becoming multilingual and the need for tools and techniques for automatic processing of languages is evident. NLP is seen as the subject dealing with such problems.

English to Telugu translator can be used in automatic translations of web. Though English is very much adopted in India, less than 5% of the population understands the language and people will understand the things better if they were told in their native language. English to Telugu translator will help people understand the works written in English in a better way. These translators will be very helpful for communication among people and for learning English.

In any translation, whether human or automated, the meaning of a text in the source language including the context must be completely translated to its equivalent meaning in the target language's translation. It is not a straight forward deal as it appears. According to Claude Bedard

“Text has its own organization and is filled with pointers that relate sentences and words into a broader picture, and that a proper translation should respect this fact” [1]. Discourse Translation is never a mere word-for-word substitution. Discourses are texts above sentence level. When a pronoun in the second sentence is referring to a subject in the first sentence it cannot be translated as a separate sentence instead while translation the first and second sentences should be interpreted as a whole and not as individual sentences[2]. The translation process involves identifying the antecedents of the anaphors which is resolution and creating the references over the discourse entity which is termed as generation [3, 4]. After resolving the anaphors the next step is to map the verbs to agree with the GNP features of the anaphors.

Discourse oriented MT makes the translations more natural in MT systems. Paragraph-by-paragraph MT seems to be a complicated task for practical needs. It involves the complete understanding of the paragraph, the determination of discourse topics, goals, intentions, so that the output can be produced in accordance with the respective discourse rules and purposes [5].

A discourse machine translation system performs a series of steps like tokenization, POS (parts of speech) tagging, parsing, reordering, reference resolution and finally, verb suffix mapping to achieve meaningful translations preserving the context. In this paper we discuss the importance of verb suffix mapping.

## 2. VERB SUFFIXES IN TELUGU AND THEIR VARIATIONS

A verb expresses an action or state of being. Telugu verbs are formed by combining roots with other grammatical information. Simple verbs in their finite forms are inflected for tense followed by GNP endings or states. In order to indicate aspect and modality of verbs various auxiliaries are employed

Ex: SL: Sita is singing

TL: సీత పాట పడుచున్నది

(Transliteration of Telugu script to English): sIta pATa pADuchunnadi

Verbs in Telugu are inflected for gender, number, person and tense. The structure of a verb is given in Fig 1.

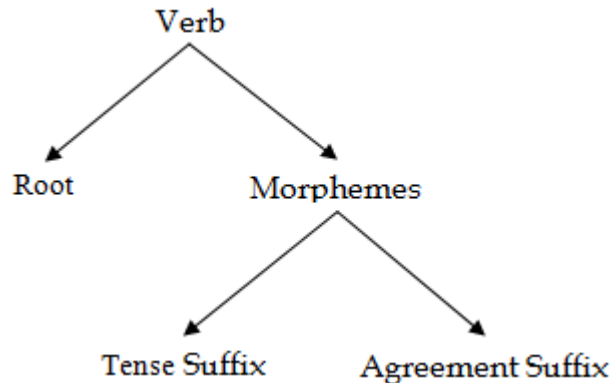


Fig. 1: Structure of Telugu Verb

pADuchunnADu is a verb whose root verb is 'pADu'. 'chunnADu' is a suffix added to the main verb to indicate the tense and feature agreement. 'chunnA' indicates the present tense and 'Du' indicates the GNP as Male, singular and third person. The structure of the verb 'pADuchunnADu' is shown in Fig 2. Similarly pADuchunnadi is a verb which inflects for tense and GNP using the suffixes, 'chunnA' and 'di', Fig 3.

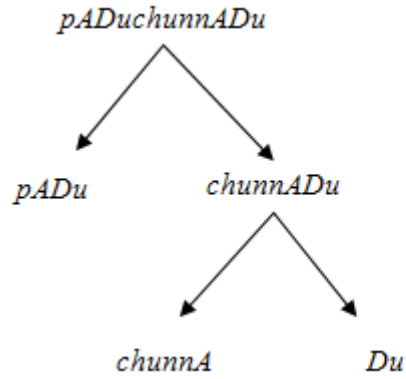


Fig. 2: pADuchunnADu

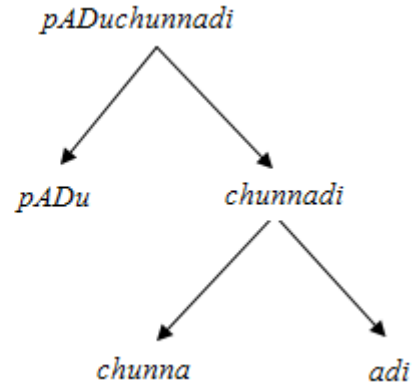


Fig. 3: pADuchunnadi

## 2.1 Subject Verb Agreement in English and Telugu

Subject-verb agreement is found in many languages, yet the degree of agreement varies considerably. In English, correspondence of a verb with its subject lies in person and number features. In Telugu language, the verb agrees with subject in gender, number and person i.e. when the subject is a third person personal pronoun, the verb agrees with subject in gender number person for remaining pronouns as subjects the verb agrees with them in person and number features only. Telugu language has got its own set of agreement rules.

- In Telugu language a finite verb exhibits agreement with nominative form of a noun in gender, number and person

Ex: SL: Boy is singing

TL: అమ్మాయి పాడుచున్నది.

Transliteration: ammAyi pADuchunnadi.

SL: Girl is singing

TL: అబ్బాయి పాడుచున్నాడు.

Transliteration: abbAyi pADuchunnADu

- All non-pronominal subjects are considered to be in 3rd person. When subject is in third person and singular in number, both feminine and neuter genders have same verb suffixes

and for masculine gender verb suffix would be different. When subject is in third person and plural in number, both masculine and feminine genders have same verb suffixes but for neuter gender the suffixes of verbs differ as shown in Table 1.

Table 1: Change of Verb suffixes with GNP features of Subject

English	Telugu	Gender	Number	Verb Suffix
Girl is singing	అమ్మాయి పాడుతున్నది	Girl, F	S	అది
Radio is singing	రేడియో పాడుతున్నది	Radio, N	S	అది
Boy is singing	అబ్బాయి పాడుతున్నాడు	Boy, M	S	ఆడు
Boys are eating	అబ్బాయిలు పాడుతున్నారు	Boys, M	P	ఆరు
Girls are eating	అమ్మాయిలు తింటున్నాయి.	Girls, F	P	ఆరు
Dogs are eating	కుక్కలు తింటున్నాయి	Dogs, N	P	ఆయి

### 3. VERB SUFFIX MAPPING

Agreement of gender number person (GNP) is realized in two cases in subject verb agreement and agreement of anaphoric pronoun with its antecedent [6]. After Anaphora generation next step is verb suffix mapping. If the anaphora is at subject position, the verb of that sentence should agree with the anaphora. In Telugu language verb inflects for gender, number and person. Subject-Verb agreement rules of Telugu are discussed in detail in section 1. After anaphora generation, grammatical gender and number information of the pronoun are required for verb suffix change.

Table 2: Third person Pronouns marking gender in English and Telugu

Pronouns in English	Mark gender	Pronouns in Telugu	Mark gender
He	Yes	అతడు	Yes
She	Yes	ఆమె	Yes
It	Yes	అది	Yes
They	No	వారు, అవి	Yes

### 3.1 Verb Dependency on Anaphors

English verbs are not strongly inflected. The only inflected forms are third person singular simple present in –s, a simple past form, a past participle form, a present participle and gerund form in -ing. Most verbs inflect in a simple regular fashion. There are some irregular verbs with irregular past and past particle forms. If pronoun is the subject then the auxiliary verb should agree with the number and person features of the subject.

Telugu verbs are formed by combining roots with other grammatical information. Simple verbs in their finite forms are inflected for tense followed by GNP endings or states. In order to indicate aspect and modality of verbs various auxiliaries are employed [7].

The structure of the verb will be like Verb stem+ Tense Suffix+ GNP Suffix. When a pronoun is the subject of a sentence, the verbs agrees in person, number, and when using third person agrees with gender also [8].

The verb inflections should agree with gender and number features of the subject, noun. Though Telugu nouns have three genders and two numbers the verb suffixes change in a different way. In singular number, feminine and neuter nouns have the same verb suffixes but masculine nouns have different verb suffixes. In plural numbers masculine and feminine nouns have same GNP endings, but for neuter nouns they differ. The suffixes for the verb ‘go’ are shown in the Table 3.

Table 3: Suffixes of verb ‘go’ for different GNP features

Person	Singular		Plural	
	Pronoun	Verb (go/goes)	Pronoun	Verb (go)
1	I (నేను) (M/F/N)	వెళ్ళాను	We (మేము) (M/F/N)	వెళ్ళాము
2	You (నీవు) (M/F/N)	వెళ్ళవు	You (మీరు) (M/F/N)	వెళ్ళారు
3	He (అతడు)(M)	వెళ్ళాడు	They (వారు)(M/F)	వెళ్ళారు
	She (ఆమె) (F)	వెళ్ళింది	They (అవి) (N)	వెళ్ళాయి
	It (అది) (N)	వెళ్ళింది		

### 3.2 Verb Patterns

Basic verb phrase patterns in English and their corresponding Telugu translations were shown in table 4. It can be noticed that any verb phrase in Telugu will end with VBD (Past tense)/ VBZ (3<sup>rd</sup> person singular)/ VBP (non 3<sup>rd</sup> person singular)/ MD (Modal)/ have/has/ had/ am/ is/ are/ was/ were, Table 4. Depending on the GNP features of the anaphor the last word of a verb phrase should change its suffix.

Table 4.9: Verb Patterns in English and Telugu

English Pattern	Telugu Pattern	Example English	Telugu Translation
Single word verbs			
VBZ	VBZ	He goes.	అతడు వెళ్ళును.
VBP	VBP	We see.	మేము చూశాము.
VBD	VBD	I left.	నేను వెళ్ళితిని.
Two word verb Phrases			
MD+VB	VB+MD	I will stay.	నేను ఉండ గలను.
have/has/had+VBN	VBN+ have/has/had	I have gone. She has gone.	నేను వెళ్ళి ఉన్నాను. ఆమె వెళ్ళి ఉన్నది.
am +VBG	VBG + am	I am going	నేను వెళ్ళుచూ ఉన్నాను.
is/are +VBG	VBG + is/are	They are going	వారు వెళ్ళుచూ ఉన్నారు.
was/were +VBG	VBG+ was/were	He was going.	అతడు వెళ్ళుచూ ఉండినాడు.
am+ VBN	VBN + am	I am done	నేను చేసినాను.
is/are +VBN	VBN+ is/are	He is released	అతడు విడుదల చేయబడి ఉన్నాడు.
was/were +VBN	VBN + was/were	She was forgiven.	ఆమె క్షమించబడి ఉండినది.
Three word Verb Phrases			
MD+have+VBN	VBN + have + MD	I could have danced	నేను అడి ఉండ గలను.
MD+be+VBG	VBG+ be+ MD	She should be arriving	ఆమె వచ్చుచూ ఉండ వలెను.
MD+be+VBN	VBN+ be + MD	He must be stopped	అతడు ఆగి ఉండ వలెను.

English Pattern	Telugu Pattern	Example English	Telugu Translation
have+been+VBG	VBG + been + have	We have been travelling	మేము ప్రయాణము చేయుచూ ఉండి ఉన్నాము.
has+been+VBG	VBG + been + has	She has been travelling	ఆమె ప్రయాణము చేయుచూ ఉండి ఉన్నది.
had+been+VBG	VBG + been + had	It had been raining	ఇక్కడ వర్షించుచూ ఉండి ఉండగలదు.
have+been+VBN	VBN+ been + have	I have been waited	నేను నిరీక్షిస్తూ ఉండి ఉన్నాను.
has+been+VBN	VBN+ been + has	She has been tortured	ఆమె వేధించబడి ఉండిఉన్నది.
had+been+VBN	VBN+ been + had	He had been tortured	అతడు వేధించబడి ఉండి ఉండ గలడు.
am+being+VBG	VBN+ being+ am	I am being groomed	నేను లాల్చించబడి ఉంటూ ఉన్నాను.
is/are+being+VBG	VBN+ being+ is/are	It is being discussed	అది తర్కించబడి ఉంటూ ఉన్నది.
was/were+being+V BG	VBN+being+was/w ere	They were being interrogated	వారు ప్రశ్నించబడి ఉంటూ ఉండిరి.
Four word verb phrases			
MD+have+been+V BG	VBG+ been+have+MD	It should have been raining	అక్కడ వర్షించుచూ ఉండి ఉండ వలెను.
MD+have+been+V BN	VBN+been+have+ MD	It should have been rained	అక్కడ వర్షించబడి ఉండి ఉండ వలెను.
MD+be+being+VB N	VBN+being+be+M D	It may be being discussed.	అది తర్కించబడి ఉండి ఉండ గలదు.

#### 4. VERB SUFFIX DEPENDENCY ON GNP FEATURES OF NOUN

Change the verb suffix according to the GNP features of a Noun and corresponding Pronoun.

Ex:1 SL: Students came to the zoo. They are watching birds.

TL: పిల్లలు జంతు ప్రదర్శన శాల కి వచ్చిరి. వారు పక్షులను చూచు చున్నారు.

pillalu ja.mtu pradarshana shAla ki vachchiri. vAru pakshulanu chUchu chunnAru.

Ex:2 SL: Monkeys are in the zoo. They are doing mischief

TL: కోతులు జంతు ప్రదర్శన శాల లో ఉన్నవి. అవి అల్లరి చేయు చున్నవి.

kOtulu ja.mtu pradarshana shAla lo unnavi. avi allari cheyu chunnavi.

Ex:3 SL: AC is not working properly. It is making loud noise.

TL: ఏసీ పని చెయ్యట లేదు. అది గట్టిగా చప్పుడు చేయు చున్నది.

EsI pani cheyyuTa lEdu. adi gaTTigA chappuDu chEyu chunnadi.

Ex:4 SL: AC is not working properly. Can the engineer repair it?

TL: ఏసీ పని చెయ్యడము లేదు. ఇంజనీరు దాన్ని బాగు చేయ గలడా?

EsI pani cheyyaDamu lEdu. i.mjanIru dAnni bAgu chEya galaDA?

In example 1 ‘they’ refers to ‘students’. The GNP features of students being (M/F, P, 3), ‘they’ is translated as ‘వారు’ and accordingly verb ‘are’ is translated as ‘చున్నారు’. In example 2 ‘they’ refers to monkeys. The GNP features of monkeys being (N, P, 3) ‘they’ is translated as ‘అవి’ and accordingly verb ‘are’ is translated as ‘చున్నవి’.

In examples 3 and 4, ‘it’ is the anaphor referring to a third person, singular pronoun of neuter gender, ‘AC’. The grammatical role of ‘it’ in both examples differ. In example 3 the anaphor is at subjective position and in example 4 the anaphor is at objective position. In English language same pronoun ‘it’ will be used at both subjective and objective positions. But in Telugu language two different pronouns are used for different grammatical roles. ‘అది’ is used for subjective and ‘దాన్ని’ is used for objective role. Consequently the verbs in the two sentences are చేయు చున్నది, చేయ గలడా respectively.



## 5. CONCLUSION

Translating texts may not be a new concept but translating texts preserving the context is an area of research which has been explored very little. Generally, the translations lack the flair of SL because of the lexical and syntactical differences of the language pairs involved in the translation. Discourse oriented MT makes the translations more natural in MT systems. The present paper discusses the importance of verb suffix mapping in the Anaphora resolutions and generation from English to Telugu language. The concept can be applied to many of the foreign languages.

## REFERENCES

- [1] Claude Bedard (2008), "Suddenly its Discourse Analysis", Language Technology 13, May-June 1989
- [2] T. Suryakanthi, Kamlesh Sharma (2015) "Discourse Translation from English to Telugu" In Proceedings of the Third International Symposium on Women in Computing and Informatics (WCI-15), ACM publishers
- [3] Jes'us Peral, Antonio Ferr'andez (2003) "Translation of Pronominal Anaphora between English and Spanish: Discrepancies and Evaluation" Journal of Artificial Intelligence Research Vol.18 pp. 117-147
- [4] T. Suryakanthi, Dr. S.V.A.V Prasad, Dr. T.V Prasad, "Translation of Pronominal Anaphora from English to Telugu Language", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 4, 2013
- [5] Hauenschild C., (1988) Discourse structure - some implications for Machine Translation, Proc. of Conf. on New Directions in Machine Translation, Budapest, August 18-19 Dodrecht-Holland
- [6] Abdel-Aal Attia Mohammed, (2002) "Implications of the Agreement Features in Machine Translation", M.A Thesis, Faculty of Languages and Translation, Al-Azhar Univ.
- [7] Krishnamurti, Bh., (1985) A Grammar of Modern Telugu, Oxford Univ. Press, New York.
- [8] Henry Arden Albert, (1905) A Progressive Grammar of the Telugu Language With Copious Examples And Exercises, S.P.C.K Press, India.

## AUTHORS

**Dr. S. Tangirala** earned her master's degree in computer applications in 2006 from Andhra University, Visakhapatnam, India and doctoral degree in 2014 from Lingaya's University, Faridabad, India. She has worked for around 2 years in software industry and has been teaching for 5 years at University Level. She was Assistant Professor of Computer Applications at Lingaya's University and worked as Fellow at Botho University, Gaborone, Botswana. Currently she is working with University of Botswana. She has 17 research papers to her credit in various international conferences and journals. Her current research interests include Artificial Intelligence, Natural Language Processing, Machine Translation, Big data analytics and Theory of automata.



## **AUTHOR INDEX**

*Abdelwahab Bourai 01*  
*Amna Al Shamsi 57*  
*Ankit Gautam 109*  
*ArchitYajnik 35*  
*Aysa Al Nuaimi 57*  
*Aysa Al Shamsi 57*  
*Elarbi Badidi 57*  
*Esraa Alshammari 121*  
*Farah Abu-Dabaseh 121*  
*Fatima Al-Raisi 01*  
*Hamida Amdouni 91*  
*Imen Mguiris 91*  
*Jenq-Haur Wang 41*  
*Matthew Purver 15*  
*Mohamed Mohsen Gammoudi 91*  
*Mohammed O.AL-AMR 131*  
*Ravi Bhushan Mishra 109*  
*Rishi Yadav 109*  
*Shahd Alharbi 15*  
*Shannon Heh 65*  
*Shaukat Ali Shahee 77*  
*Suryakanthi Tangirala 143*  
*Usha Ananthakumar 77*  
*Weijian Lin 01*  
*Zewdie Mossie 41*