Taking innovative machine learning research to industry

Damminda Alahakoon

SAS Analytics innovation Lab
Research Centre for Data Analytics and Cognition
La Trobe University
Presenter Bio

- **Damminda Alahakoon**
- Professor and Director, Research Centre for Data Analytics and Cognition
- La Trobe University, Melbourne, Australia

- Over 10 years’ experience in the IT and finance industries, as a credit officer and Data Mining Specialist, in Sri Lanka, Australia and the Netherlands.

- 12 years as an academic

- Key research expertise in the areas of Data Mining, Predictive Analytics, Text Analytics, Machine Learning and Business Intelligence.

- Published over 100 research articles in data mining, clustering, neural networks, machine learning and cognitive systems.
Melbourne, Australia

2017 Most liveable cities in the world

1. Melbourne, Australia
2. Vienna, Austria
3. Vancouver, Canada
4. Toronto, Canada
5. Calgary, Canada
6. Adelaide, Australia
7. Perth, Australia
8. Auckland, New Zealand
9. Helsinki, Finland
10. Hamburg, Germany
La Trobe University

Established in 1964
36,000 students
1,500 academic and
1,700 admin staff

Named after the first lieutenant-governor of the state of Victoria

Faculties: Engineering, Computing, Business, Law, Health, Arts and Social Sciences
Research Centre for Data Analytics and Cognition
SAS Analytics Innovation Lab

Foundation Research
- Machine learning & deep learning
- Artificial intelligence
- Data mining
- Cognitive Neuroscience
- Parallel computing
- Text analytics

CDAC Analytics Building Blocks
- Handling large data volumes
- Text Mining
- Adaptive, real time learning
- Stream analysis
- Data integration
- Sequence capture

Enrich and Integrate with Other Technologies
- SAS
  - Enterprise Miner
  - Visual Analytics
  - Social media suite
  - R, python etc
  - NLP
  - Other ...

Applications and Training
- Health
- Defence and national security
- Finance
- Social media
- Teaching and Learning
- Used as supporting technologies in the La Trobe Masters of Business Analytics

Projects and Collaborations
- Internships with SAS Clients
- Masters Projects
- PhD Industry projects
- Industry
- Government
- Academia
La Trobe Masters of Business Analytics

<table>
<thead>
<tr>
<th>Foundation</th>
<th>BUS5SMM Sustainable Marketing and Management</th>
<th>BUS5IAF Introduction to Accounting and Finance</th>
<th>BUS5BIM Business information management</th>
<th>BUS5SBF Statistics for business and finance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core analytics</td>
<td>BUS5PB Principles of business analytics</td>
<td>BUS5PA Predictive analytics</td>
<td>BUS5VA Visual analytics</td>
<td>BUS5WB Data warehousing and big data</td>
</tr>
<tr>
<td>Advanced analytics</td>
<td>BUS5CA Customer analytics</td>
<td>BUS5AP Analytics in practice</td>
<td>CSE5DWD Data warehouse concepts and design</td>
<td>CSE4DSS Decision support systems</td>
</tr>
<tr>
<td>Specialisation</td>
<td>4x electives, can be used to complete a specialisation, e.g., marketing analytics, sports analytics, data science, etc.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Guest lectures by SAS consultants
- Joint certification with SAS
- Internship and placement opportunities through SAS
- SAS voucher system (unique) – where students can access SAS industry training up to three years after graduation at highly discounted prices.
Data and Information in the Age of Big Data

- **Volume**
- **Volatility**
- **Variety**
- **V**...

### Media as a big data source
- Images, videos, audio, podcasts
- Social media platforms like Facebook, Twitter, YouTube, Instagram

### Cloud as a big data source
- Public, private, or third party cloud platforms

### Web as a big data source
- Data publicly available on the web

### IoT as a big data source
- Data generated from the interconnection of IoT devices

### Databases as a big data source
- Traditional and modern databases

#### Text, images, video..
- Semantics, meaning, culture etc
- Social media – sentiment, emotions, human behaviours and traits...

#### Machine and process generated
- Can relate to behaviours undefined by humans

#### High frequency of generation
- Could be hundreds of readings per second (or more)
- Patterns can exist in different levels of abstraction/aggregation

[https://www.allerin.com/blog/top-5-sources-of-big-data](https://www.allerin.com/blog/top-5-sources-of-big-data)
Difficulties for Machine Learning

What is data? What does data represent?

Traditionally

Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing - Wikipedia

Can be labelled
Even when unstructured can transform in to structured form
Suitable for Machine Learning

Big Data (New environment)

Text, images, video.. Machine and process generated Machine and process generated

May not relate to actual (known) objects or events
May represent emotions, behaviours without clear classifications
May be highly granular data points which require aggregation to be 'meaningful'

Difficult or impossible to label
Unstructured data may contain emotions and individuality which are difficult to capture

Not ideal for traditional Machine Learning
### Our Approach in New Machine Learning Algorithms

To address problems due to: Unlabelled, unstructured and highly granular data

<table>
<thead>
<tr>
<th>Unsupervised machine learning</th>
<th>Self structuring algorithms</th>
<th>Situation capture from unstructured data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Self organising maps</td>
<td>• Models which adapt structure and evolve to represent data</td>
<td>• Enrich traditional ‘data’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cognitive psychology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Emotion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Behavioural traits</td>
</tr>
</tbody>
</table>

Continuous monitoring and incremental learning algorithms
1. Self structuring and unsupervised learning

Growing SOM Algorithm

Self Organizing Map (SOM) – fixed structure

Growing SOM – self structuring

Growing SOM – Generating cluster hierarchies
1. Self structuring and unsupervised learning

Growing SOM Algorithm

- World
  - Telegraph London
  - North Korea
  - Climate change
  - Prime minister
  - Report
  - Service
  - Plan
  - Nation
  - State

- Entertainment
  - Sydney opera house
  - Guardian media
  - People
  - Show
  - NSW
  - Australia
  - Family
  - House
  - Month

- National
  - University of Sydney
  - Financial crisis
  - Supreme court
  - Police officer
  - State government
  - Federal government
  - Chief executive
  - Minister
  - Australia
  - Govern

- Sport
  - Sydney turf club
  - Sydney cup
  - Melbourne cup
  - Champion league
  - Caulfield cup
  - Run
  - Start
  - Horse
  - Final
  - Win

- Jobs, housing, income
  - acq, earning, interest, money-fix, trade, reserve, money-supply
  - acq, earning, interest, money-fix, trade, reserve, money-supply

- Interest
  - SF = 0.001

- Trade
  - SF = 0.1

- Acq
  - SF = 0.5
2. Cognitive models for situation capture

Event indexing model based base event integration
2. Cognitive models for situation capture
3. Capturing emotion and personality traits
4. Monitor behaviour over time

GSOM as a base building block to capture data movement over time (IKASL algorithm)
Generates ‘cluster pathways’

Detecting potential events using changes in data volume and sentiment (text/social media application)
Captures events in cluster pathways
## Text Mining and Sentiment Extraction Tools Used

<table>
<thead>
<tr>
<th>Text pre-processing</th>
<th>Text feature representation</th>
<th>Entity mining</th>
<th>Topic mining</th>
<th>Opinion mining</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text Filtering:</strong></td>
<td><strong>Document frequency based:</strong></td>
<td><strong>Entity recognition:</strong></td>
<td><strong>Text clustering:</strong></td>
<td><strong>Sentiment analysis:</strong></td>
</tr>
<tr>
<td>- SAS Text Filter</td>
<td>- SAS Text Parser</td>
<td>- Stanford NER</td>
<td>- SAS Text clustering</td>
<td>- SAS sentiment analysis workbench</td>
</tr>
<tr>
<td>- NLTK python library</td>
<td>- NLTK TfidfVectorizer</td>
<td>- OpenNLP NER</td>
<td>- Weka java library</td>
<td>- Stanford NLP sentiment analysis</td>
</tr>
<tr>
<td>- Text Mining (tm) R package</td>
<td>- Weka</td>
<td></td>
<td>- Sklearn python library</td>
<td>- SentiWordnet</td>
</tr>
<tr>
<td>- Weka</td>
<td>- JATE (Java Automatic Term Extraction toolkit)</td>
<td></td>
<td></td>
<td>- SentiStrength</td>
</tr>
<tr>
<td><strong>Text parsing:</strong></td>
<td><strong>Word-embedding based:</strong></td>
<td><strong>Coreference resolution:</strong></td>
<td><strong>Topic modelling:</strong></td>
<td><strong>Emotion analysis:</strong></td>
</tr>
<tr>
<td>- SAS Text Parser</td>
<td>- Stanford NLP Coreference Resolution</td>
<td>- SAS Text topic</td>
<td>- WordNet-Affect</td>
<td></td>
</tr>
<tr>
<td>- NLTK text tokenizer</td>
<td>- Word2Vec</td>
<td>- Mallet</td>
<td>- EmoLex</td>
<td></td>
</tr>
<tr>
<td>- OpenNLP sentence parser, tokenizer</td>
<td>- GloVe</td>
<td>- Gensim</td>
<td>- DeepMoji</td>
<td></td>
</tr>
<tr>
<td>- Stanford NLP sentence parser, tokenizer</td>
<td></td>
<td>- Sklearn</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example Application 1: Cluster pathway based event detection from tweets

topic pathways of #Microsoft dataset

1. Microsoft accepts Bitcoin to buy Xbox games and Windows apps.
2. Google published a Windows 8.1 vulnerability.
4. Microsoft is thinking to make Windows an open source operating system in coming years.
5. Microsoft mentioned that Windows 10 would be the "last version" of Windows.
6. Facebook and Microsoft announced a partnership on Virtual Reality project.
Example Application 2: Detecting depression from social media posts

Each sentence in the social media post is represented as feature vector using three sentence representation techniques.
1. A psycholinguistic representation that employs features delineated in LIWC.
2. A deep emotion based sentence representation which is captured using DeepMoji.
3. A word embedding based technique that uses a RNN to capture and encode relevant semantic and topical aspects of depression related discourse.

<table>
<thead>
<tr>
<th>Online Support Group</th>
<th>Depression themed</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthboards</td>
<td>23,528</td>
<td>-</td>
</tr>
<tr>
<td>Healingwell</td>
<td>14,984</td>
<td>-</td>
</tr>
<tr>
<td>Patientinfo</td>
<td>6,413</td>
<td>193,098</td>
</tr>
<tr>
<td>Beyondblue</td>
<td>2,696</td>
<td>-</td>
</tr>
<tr>
<td>DailyStrength</td>
<td>1,498</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>49,119</td>
<td>193,098</td>
</tr>
</tbody>
</table>

Top three posts with highest probability of being depression related in Pizza-request dataset

Forgive me if there’s a better place for this, but I thought I’d try here. I’m a divorced dad with limited funds. The ex is remarried and in a better financial place. Not trying to compete with her, but I hate always telling the kids we can’t go out to eat. ...

today’s been a pretty sh**y day. I’ve been suffering from depression all semester long, and it’s really affected my school work. I also had a really sh**y therapy session today that mainly resulted in a lot of crying. ...

I just moved to Indianapolis while I take a semester off of college from Indiana State and I have no food or drink. .... I’m just in a bad spot tonight so if I could get a pizza bro’d to me that would be great. ...
Example Application 2: Discussion

Comments from computer science/AI conference reviewers

Extracts from Review 1
This is also a problem since the supervision is given by the discussion topic (explicitly listed as depression), while everything not in there is supposed to be from the negative class, as if people never show signs of depression in other topics. This would need to be proven, or evaluated, especially to confirm the distribution of positive/negative instances.

Extracts from Review 2
I think the critical point in depression recognition is to determine whether or not the target person/sample is a depressed patient. The person who post on the forum may not truly suffer the depressive disorder according to a clinical criterion. In the paper, you should figure out how to ensure or filter out the samples came from true depressed individuals.

Early signs of depression are traditionally diagnosed in primary care and assessed using Depression Screening Questionnaires. However, recent studies have found that this approach only has a positive case recognition rate of 36%-56%.

Extracts from Review 2
The only representation of the three that one would use is pretrained word2vec embeddings. The conclusion is that LIWC and DeepMoji are just useless, what is the point of using them? It is a negative result, and it is not surprisingly that pretrained embedding are good at this (we already know that BiLSTM work).

<table>
<thead>
<tr>
<th>DeepMoji feature</th>
<th>Description</th>
<th>% AUC drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>😞 Pensive face</td>
<td>pensive face: involved or engaged in deep serious thought</td>
<td>6.31</td>
</tr>
<tr>
<td>😞 Confounded face</td>
<td>frustrated, confused, failed or bewildered</td>
<td>5.82</td>
</tr>
<tr>
<td>😞 Crying face</td>
<td>crying with an attitude that other person has caused it</td>
<td>4.69</td>
</tr>
<tr>
<td>🎨 Face with a medical mask</td>
<td>sick or do not want to talk</td>
<td>1.79</td>
</tr>
<tr>
<td>😊 Grinning face</td>
<td>smiling</td>
<td>1.41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LIWC feature</th>
<th>Sample terms</th>
<th>% AUC drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological: Body</td>
<td>body, face, head, heart, muscle,</td>
<td>7.34</td>
</tr>
<tr>
<td>Affective: Sadness</td>
<td>sad, depress, cry, abandon, broke</td>
<td>6.82</td>
</tr>
<tr>
<td>Biological: Health</td>
<td>addict, bleed, heal, asthma, cramp</td>
<td>2.33</td>
</tr>
<tr>
<td>Regular verbs</td>
<td>ask, brought, call, care, gave</td>
<td>2.12</td>
</tr>
<tr>
<td>Auxiliary verbs</td>
<td>are, be, can, must, will</td>
<td>2.11</td>
</tr>
</tbody>
</table>
Example Application 3: Emotion analysis of hotel reviews using deep learning and SAS

80,000 Trip Advisor Reviews for 420 Hotels in Vietnam

Emoji based probability representation using Deep Learning model

Identify emotion clusters

Hierarchy clustering

Emotion clusters associated with different aspects in a hotel

- Rooms
- Food & Beverages
- Bathroom
- Location
- Public areas
- Price
- Exterior & General
- Front Desk

Identify text clusters
Example Application 4: Emotion analysis from online cancer forums

Online forum posts

“I’m not happy about my condition now but I’m terrified about the pain after the surgery, I have heard it can be unbearable…”

Emotions: SAD, AFRAID

Plutchik’s wheel of emotions

Emotion dictionary using LIWC

Comprehensive natural language processing techniques

Case study: Different emotions detected in prostate cancer patients using online forums
Example Application 5: Analysing data from ‘free living’ accelerometry data

Self-organisation of Free-living activity, GSOM / SF=0.3

- Moderate-Vigorous Intensity Activity
- Light Intensity Activity
- Sedentary Behaviour
Thank You