Durham University

Sentiment Analysis Through the Use of Unsupervised Deep Learning

S7330

Monday 8th May 2017
GPU Technology Conference

A. Stephen McGough, Noura Al Moubayed

Durham University, UK

Intel Parallel Computing Centre

TOP 500
The List.

NVIDIA
CUDA™ RESEARCH CENTER

Outline

- **The Problem**
- Unsupervised Deep Learning Model
- Text Processing for Topic Modelling
- Detecting Anomalies in Text
- Sentiment Classification

The Problem

- “90% of all the data in the world has been generated over the last two years”… IBM

- “85% of worldwide data is held in un-structured formats”… Berry and Kogan

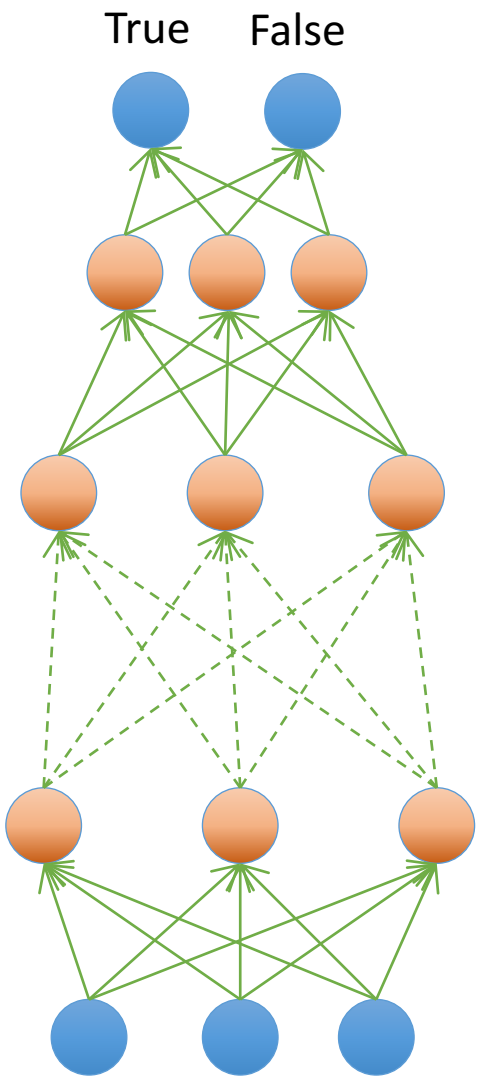
- How can we understand it? ….or better still make use of it?
- How can we determine the most pertinent information? …and then act on it?
- How can we find the needle if we are not sure what it looks like or what hay looks like?

- "Without labelling, you cannot train a machine with a new task”… IBM

Outline

- The Problem
- **Unsupervised Deep Learning Model**
- Text Processing for Topic Modelling
- Detecting Anomalies in Text
- Sentiment Classification

Deep Learning: Categorization and Regression



Output

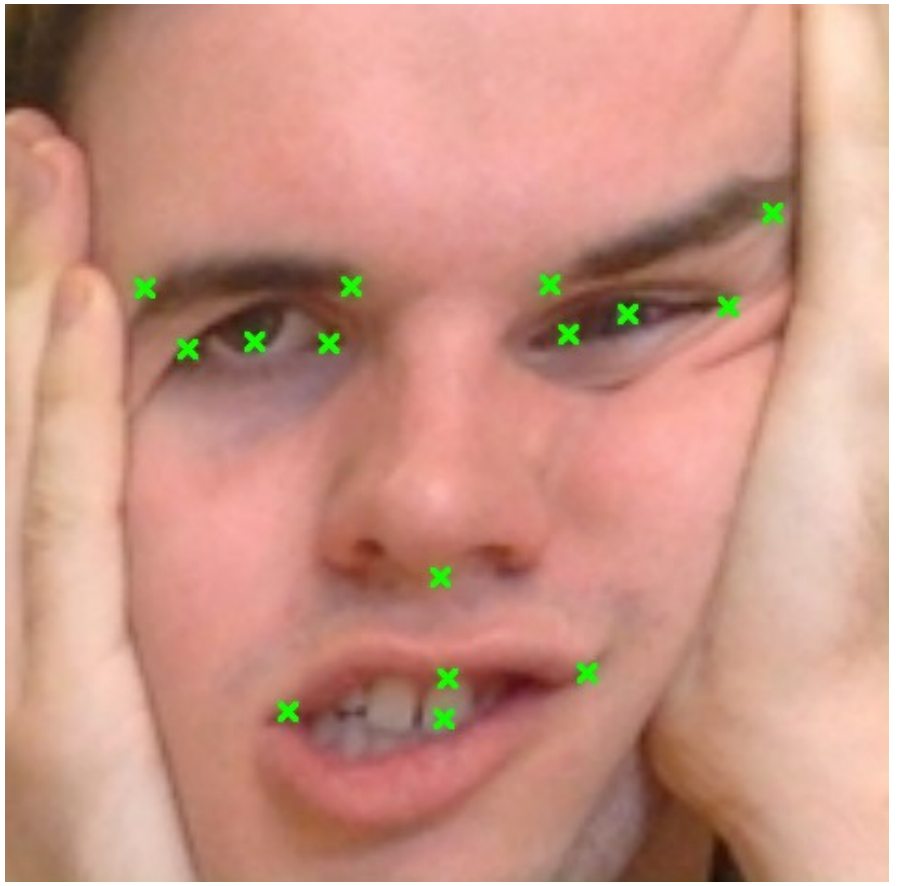
Hidden Layer

Hidden Layer

⋮

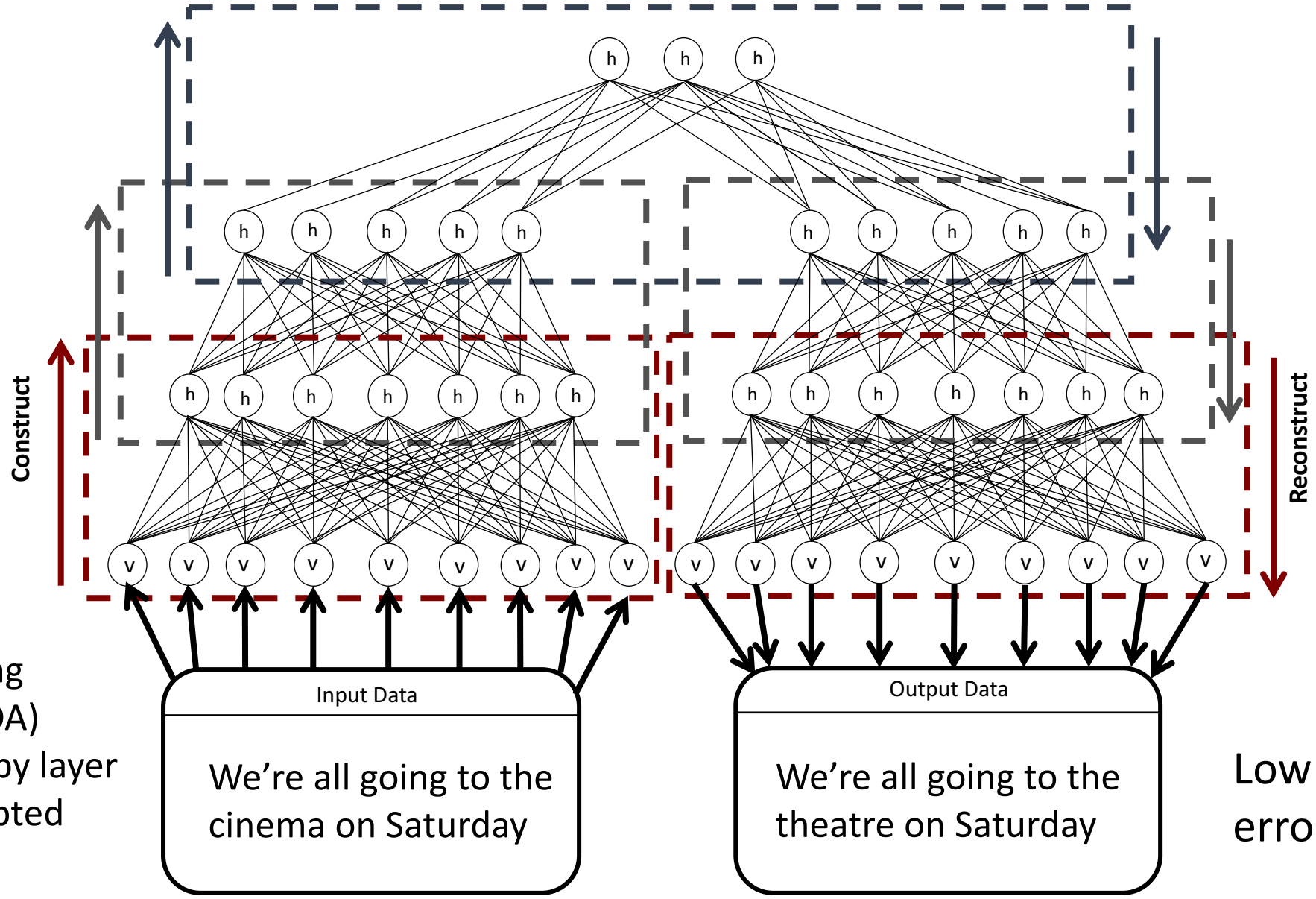
Hidden Layer

Input Vector



But these require labelled data for training

Anomaly Detection: Unsupervised Deep Learning

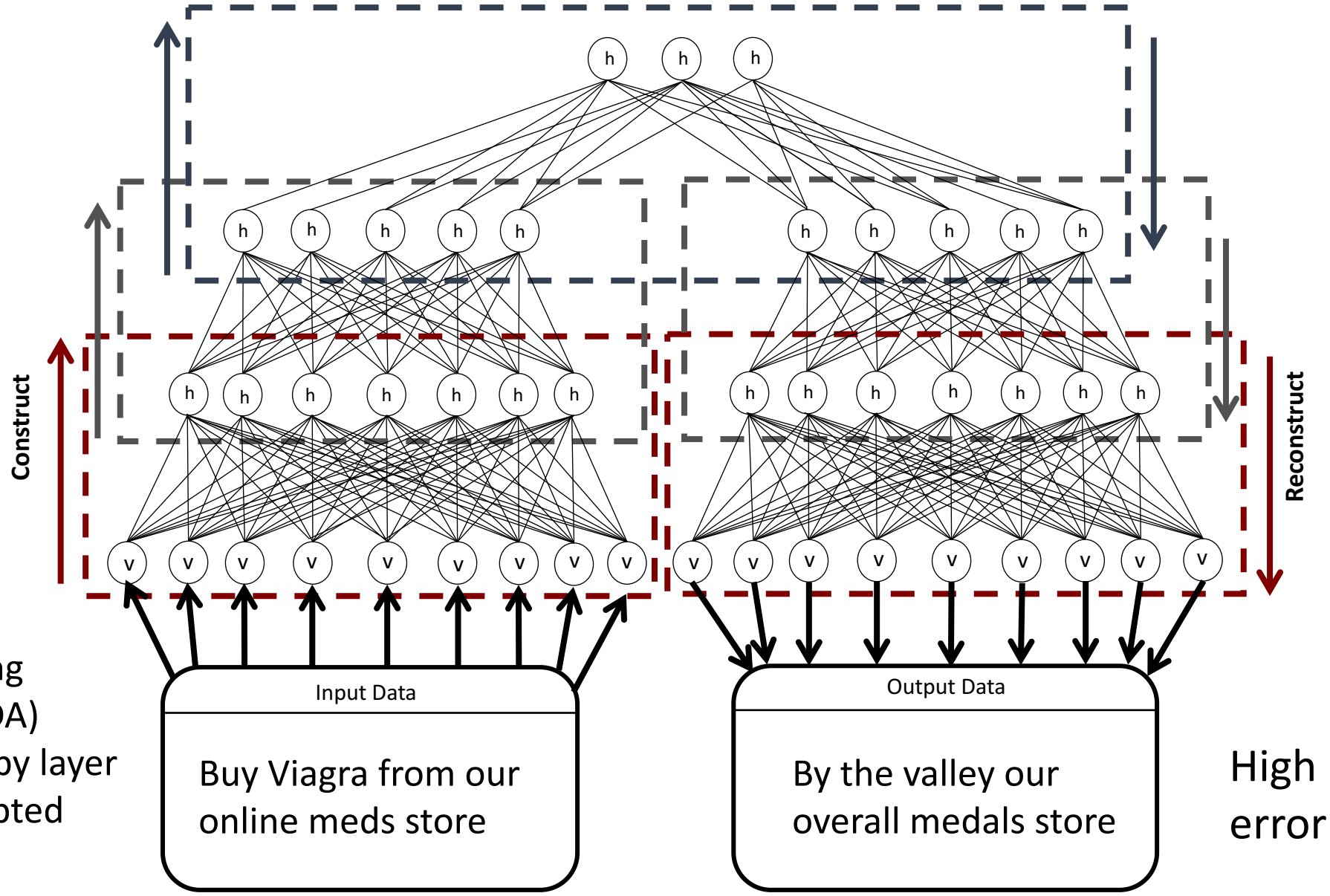


Stacked Denoising Autoencoder (SDA)

- Trained layer by layer
- Input is corrupted with noise

Low reconstruction error

Anomaly Detection: Unsupervised Deep Learning



Stacked Denoising Autoencoder (SDA)

- Trained layer by layer
- Input is corrupted with noise

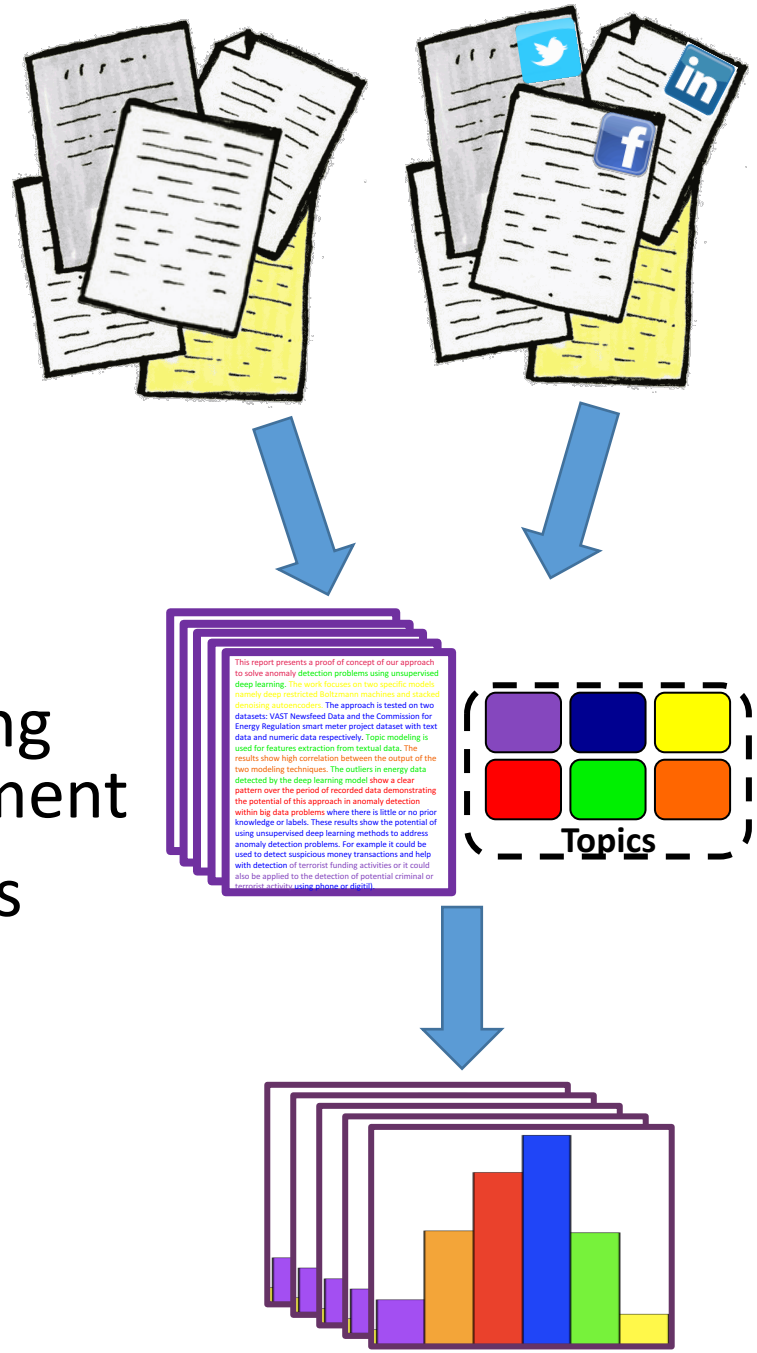
High reconstruction error

Outline

- The Problem
- Unsupervised Deep Learning Model
- **Text Processing for Topic Modelling**
- Detecting Anomalies in Text
- Sentiment Classification

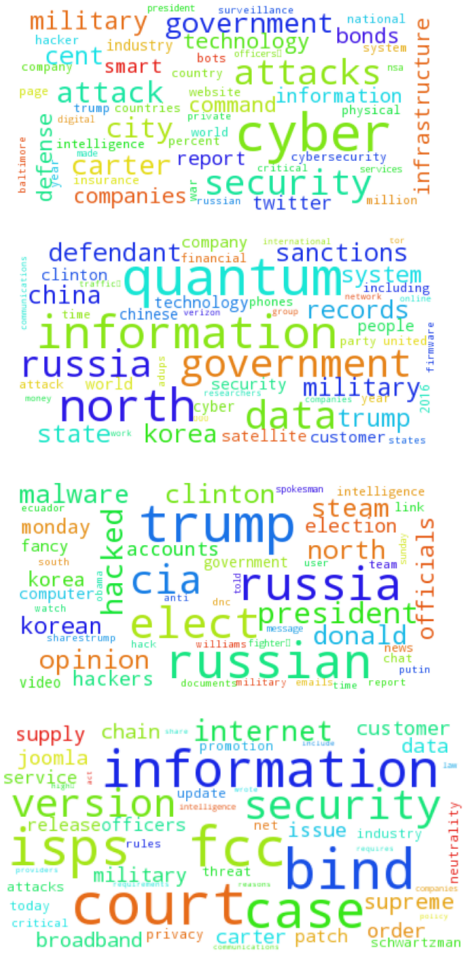
Probabilistic Topic Modelling

- Unsupervised analysis of text
  - Too many documents to label manually
- Allows us to uncover automatically themes that are latent in a collection of documents
- Same words may have different meanings depending on their co-occurrence with other words in a document
- Statistically identify the topics in a set of documents
  - Which topics appear in which documents
- Statistically identify words in topics
  - Which words co-occur in a document containing topic X
- Document transformed to feature vector of topics

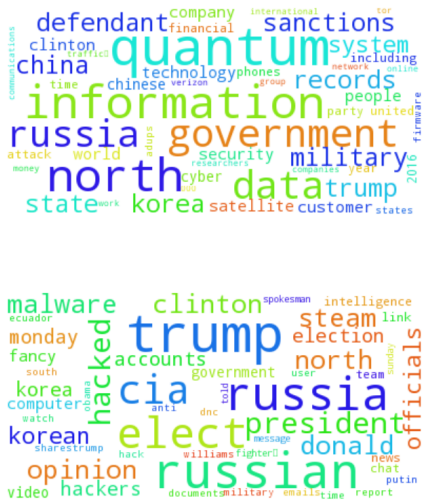


Topic Modelling:
Latent Dirichlet Allocation (LDA)

Topics



Randomly Select topics



Randomly Select words from topics

US Govt Data Shows Russia Used Outdated Ukrainian PHP Malware

The United States government earlier this year officially accused Russia of interfering with the US elections. Earlier this year on October 7th, the Department of Homeland Security and the Office of the Director of National Intelligence

Topic Modelling: Latent Dirichlet Allocation (LDA)

- Each document contains a proportion (**θ**) of each topic
  - Proportion may be zero
- The probability of each word appearing in a topic (**β**)
- LDA assumes documents were constructed via:

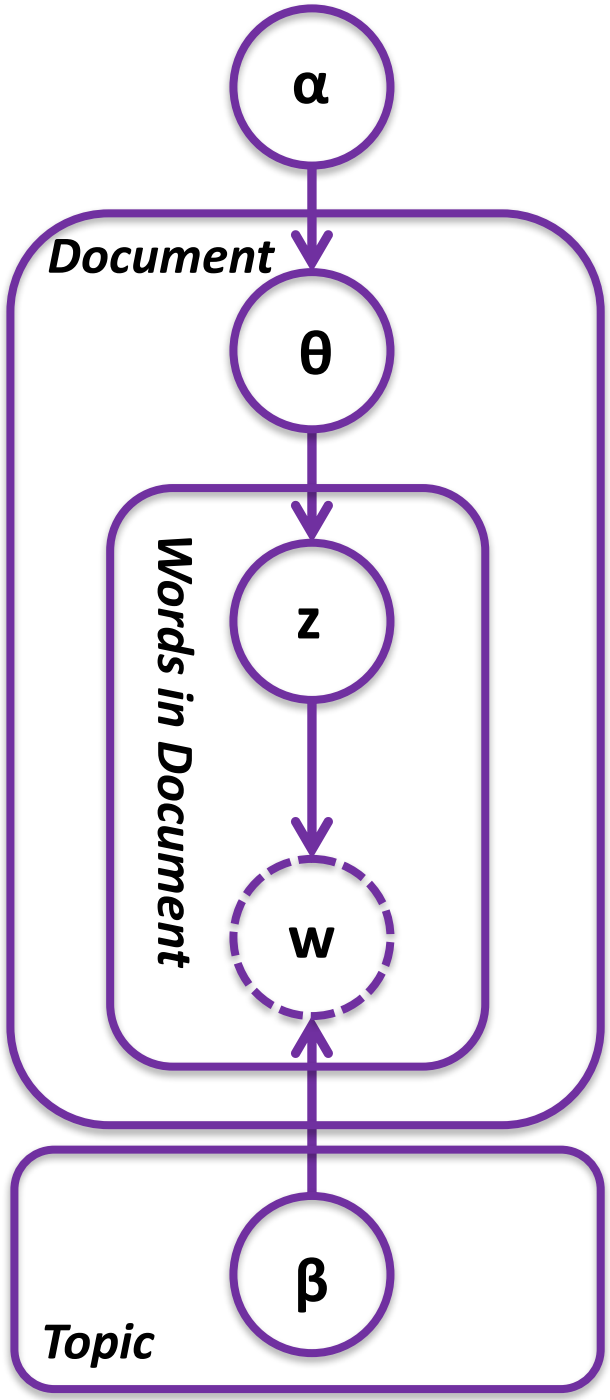
Words in document = N

Per document topic proportions = **θ** = Dirichlet function (**α**)

For each of the N words w_n:

- Choose a topic z_n ~ Multinomial(**θ**)
- Choose a word w_n from topic z_n (based on probabilities **β**)

- Need to determine **α, β, θ** given we have the documents
  - Solve using variational Bayesian methods or Gibbs sampling



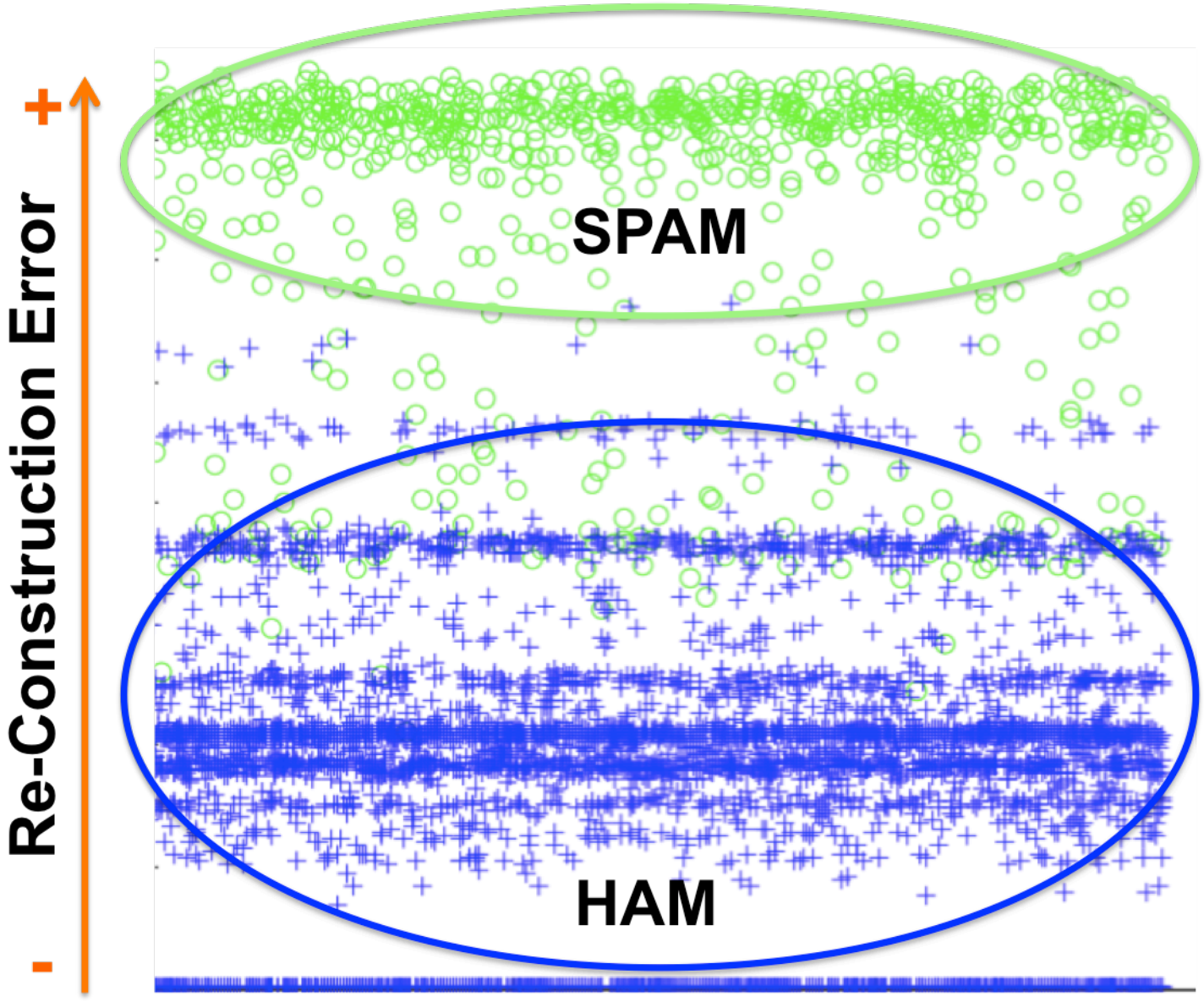
Outline

- The Problem
- Unsupervised Deep Learning Model
- Text Processing for Topic Modelling
- **Detecting Anomalies in Text**
- Sentiment Classification

Example: Anomaly Identification

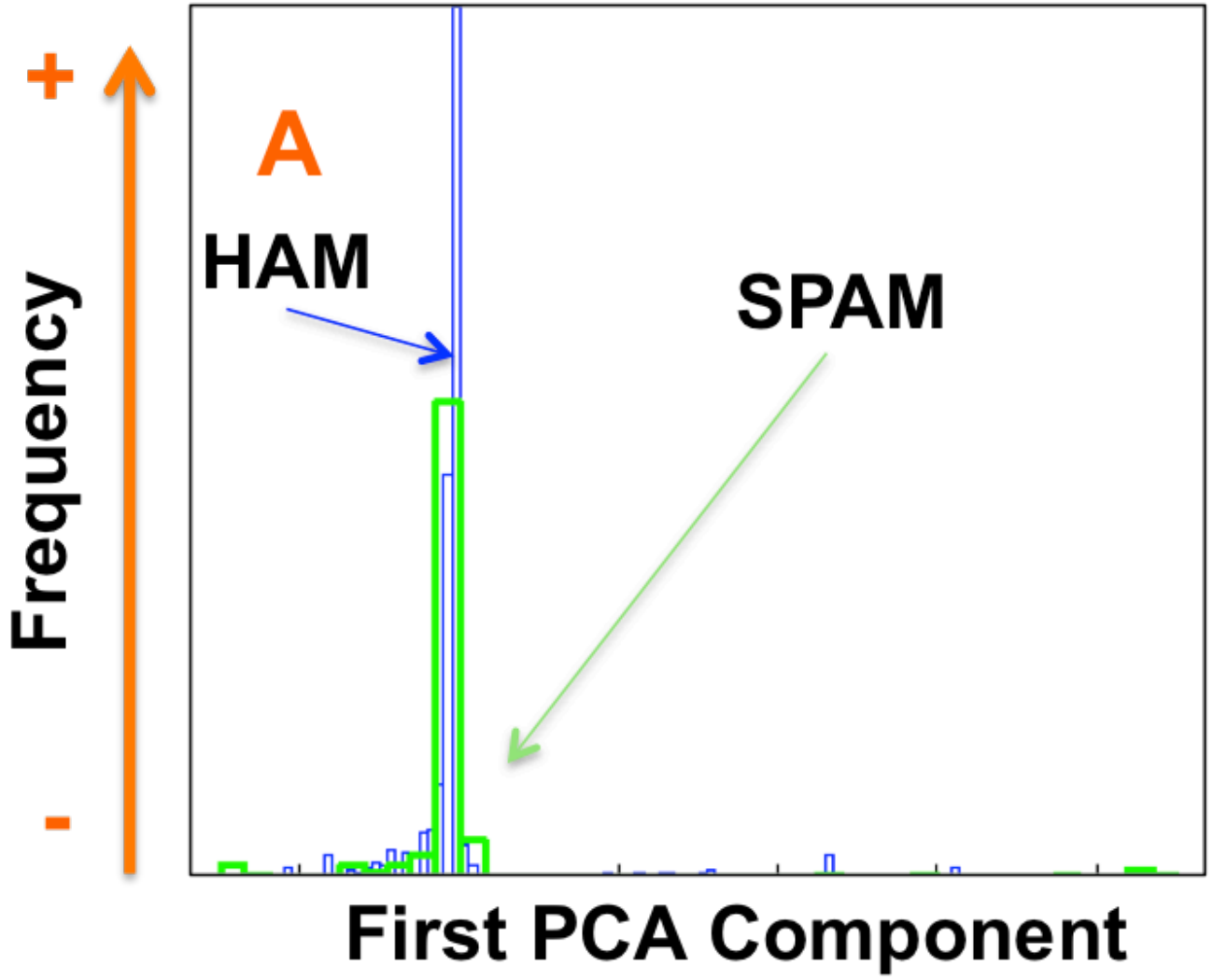
SPAM and HAM in SMS

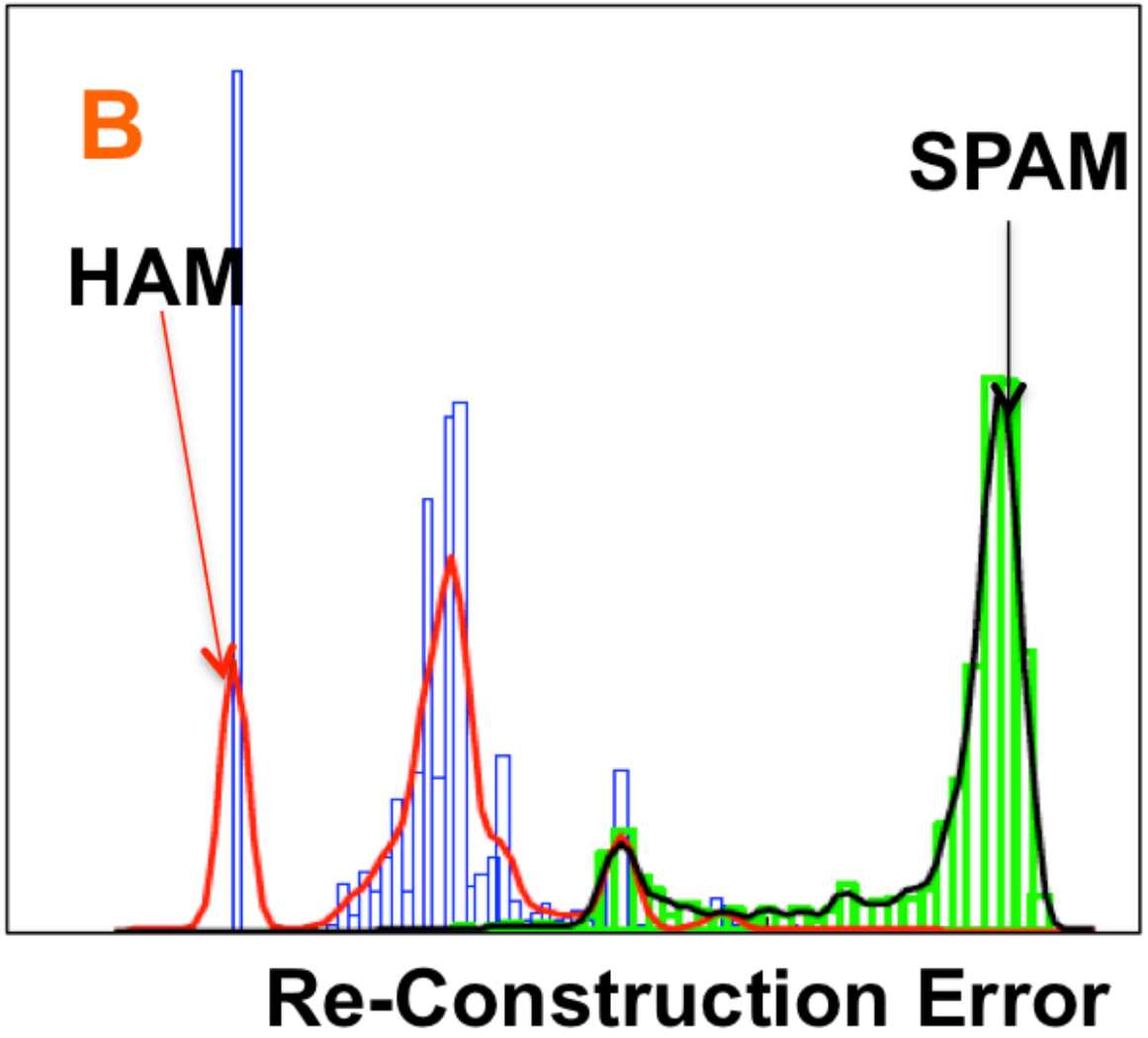
- Train LDA using the labelled data only e.g. SPAM
- Unlabeled (HAM+SPAM) used to train SDA
- Auto-identification of SPAM from HAM in SMS messages



Example: Anomaly Identification

SPAM and HAM in SMS



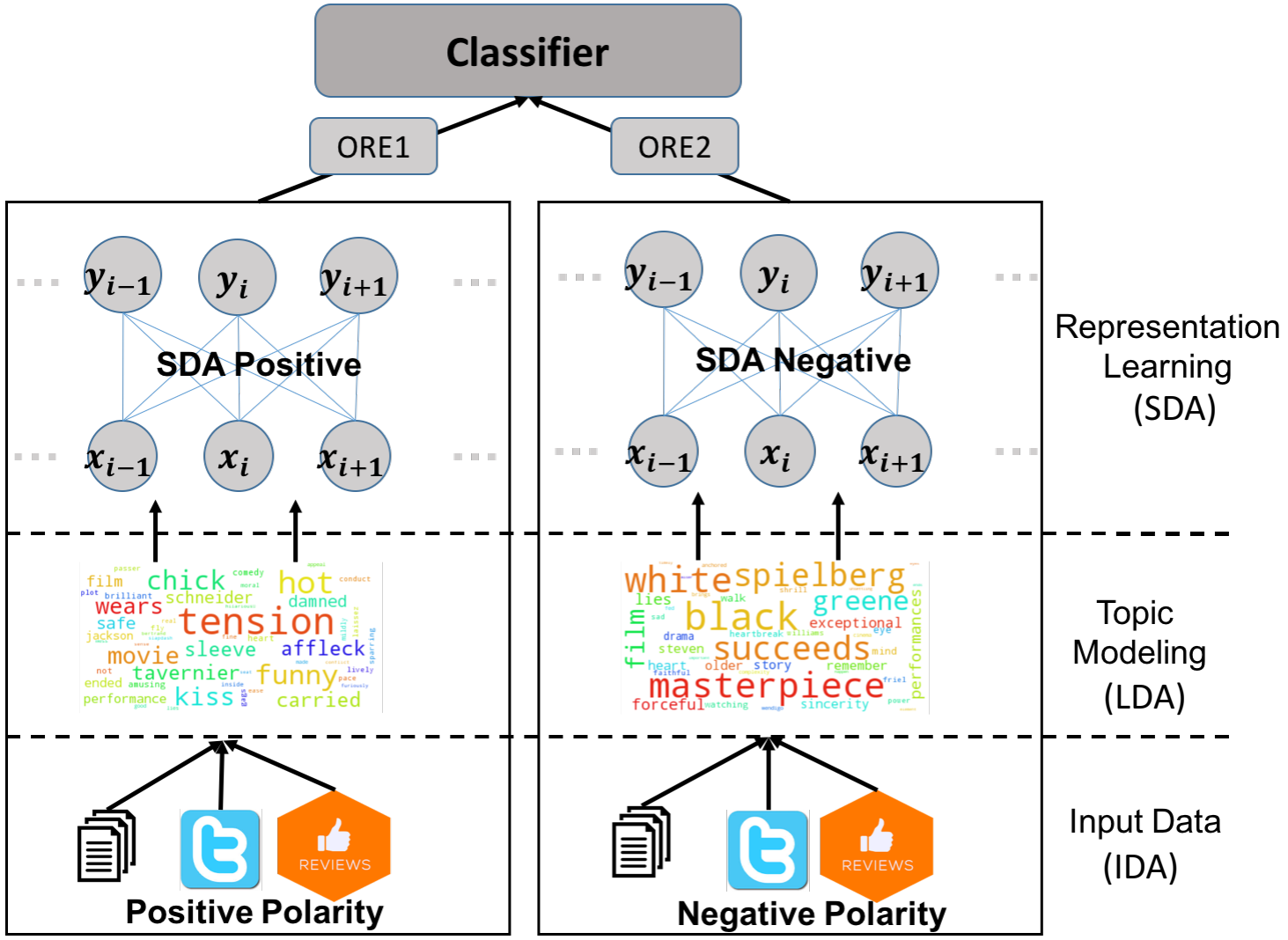


Outline

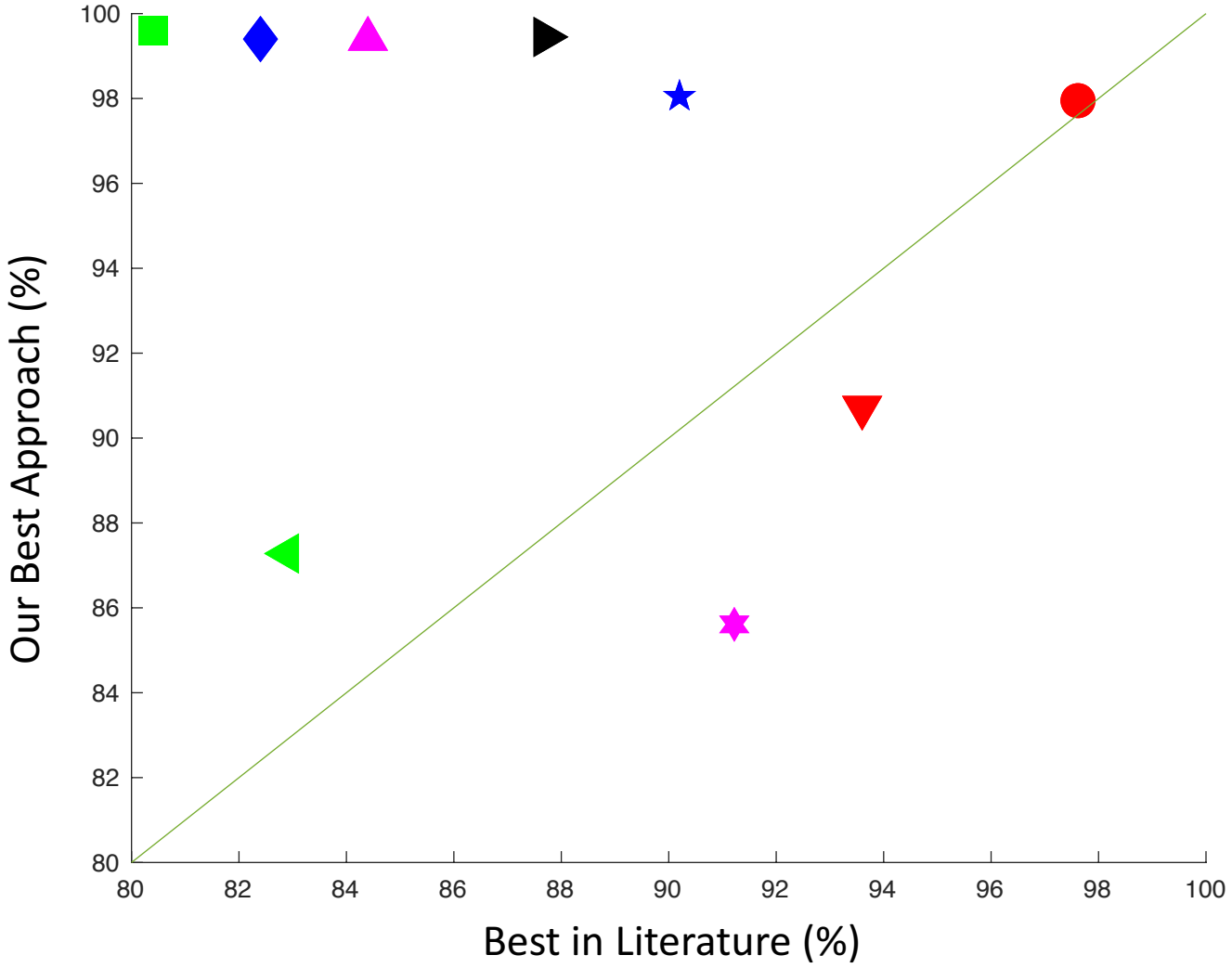
- The Problem
- Unsupervised Deep Learning Model
- Text Processing for Topic Modelling
- Detecting Anomalies in Text
- **Sentiment Classification**

Multiple Labels Example: Deep Learning for Sentiment Classification

- Train an anomaly detector on each polarity of sentiment
  - Can have many
  - E.g. positive & negative view
- Generate overall reconstruction error (ORE) for each label
- Use simple classifier on OREs to identify class

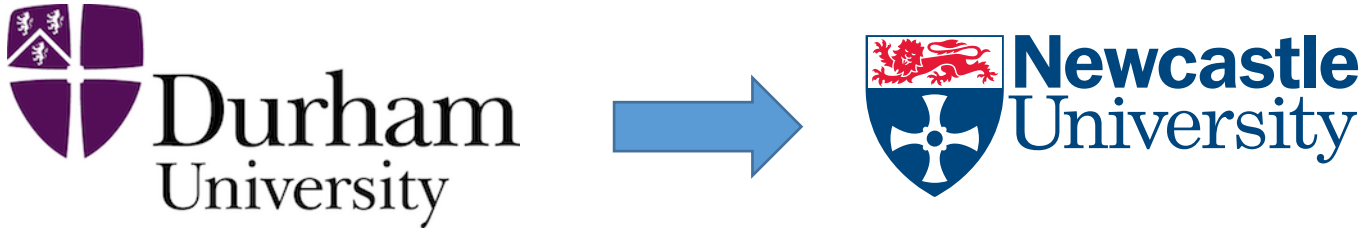


Performance Analysis



Problem	Best Literature	Best Our Approach
SMS Spam	97.64	**97.92**
MDSD-B(%)	80.4	**99.6**
MDSD-D(%)	82.4	**99.4**
MDSD-E(%)	84.4	**99.4**
MDSD-K(%)	87.7	**99.45**
Movie-Sub	**93.6**	90.72
Movie-Rev1	82.9	**87.28**
Movie-Rev2	90.2	**98.05**
IMDB	**91.22**	85.61

Summary



- Stacked Denoising Autoencoder allow us to spot anomalies within unlabeled data
- Probabilistic Topic Modelling allows us to look at documents at a higher level than just words
- When we have labels we can train on the labels to:
  - Identify sentiment
  - Classify the data

We Are recruiting:
- 2 PostDoc (Machine Learning / NLP)
- 1 PostDoc (Parallel Programming)
- Always looking for good PhD Candidates

stephen.mcgough@newcastle.ac.uk
noura.al-moubayed@dur.ac.uk