APPLICATION OF NATURAL LANGUAGE PROCESSING AND DEEP LEARNING APPROACHES TO NATURAL LANGUAGE CONTAINED IN SEC FILING DATA

Tornike Tsereteli

Computational Linguist, Reportix

Dennis Knochenwefel

CEO, Reportix

OUTLINE

1. Introduction: What is …
   - Artificial Intelligence (AI)
   - Machine Learning (ML)
   - Deep Learning (DL)
   - Natural Language Processing (NLP)?
2. The Feasibility Study
3. Results & Evaluation
4. Problems & Obstacles
5. Outlook

INTRODUCTION

ARTIFICIAL INTELLIGENCE (AI)

- Intelligent machine
- Goal: solve problems (better than/like) a human
- Solutions:
  - Autonomous vehicles
  - Financial services
  - Generating art
  - Image recognition
  - Medical diagnosis
  - Natural language processing
  - Personal assistants
  - Playing games

MACHINE LEARNING (ML)

- Feeding an algorithm data in order to make intelligent decisions in new situations
- Pros:
  - Easy to interpret results
  - Works well on small datasets
  - Computationally (financially) inexpensive
- Cons:
  - Manual feature (characteristic) engineering = time consuming & expert knowledge required
  - Bad performance on unseen situations

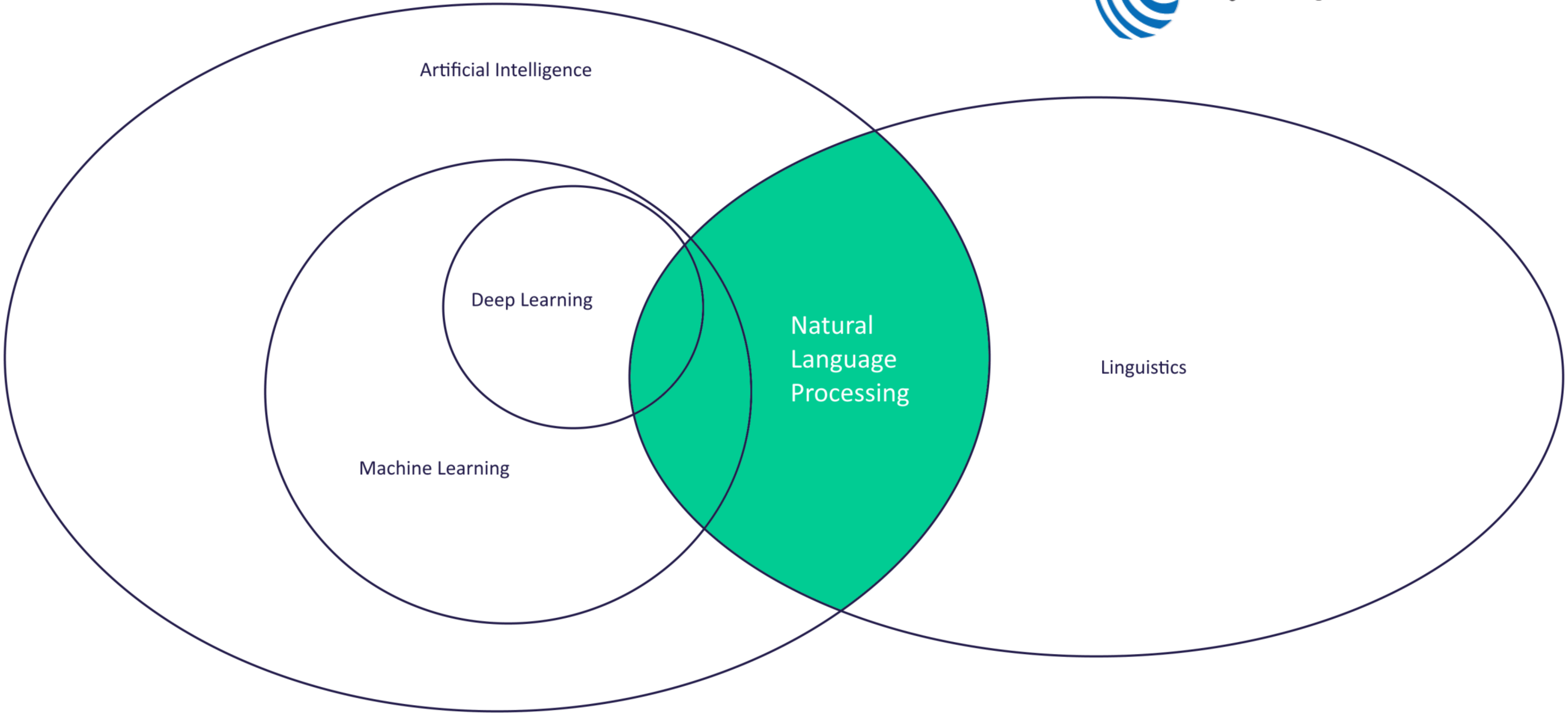
DEEP LEARNING (DL)

- "Deep" neural networks (algorithms) make intelligent decisions in new situations & on new domains
- Pros:
  - State-of-the-art performance
  - Scales well: more data (usually) = better performance
  - No feature engineering (networks learn features independently)
  - Good performance in new situations
- Cons:
  - Computationally (financially) expensive
  - Hyperparameter tuning (time consuming)
  - Black-box

NATURAL LANGUAGE PROCESSING (NLP)

- NLP is concerned with the interaction between computers and human (natural) languages
- Examples:
  - "I heard the music in my room" - Relationship Extraction
  - "The cat ate the mouse. It was small." - Coreference Resolution
  - "I'd recommend the product to anyone who loves wasting money." Sentiment Analysis
- NLP tasks:
  - Automatic Summarization, Coreference Resolution, Machine Translation, Natural Language Generation/Inference/Understanding, Named Entity Recognition, Relationship Extraction, Question Answering, Sentiment Analysis, Speech Recognition, Text-to-speech

AI & LINGUISTICS



What's so deep about Deep Learning?
What's the difference to Machine Learning?

Deep neural networks

DEEP NEURAL NETWORK (DNN)

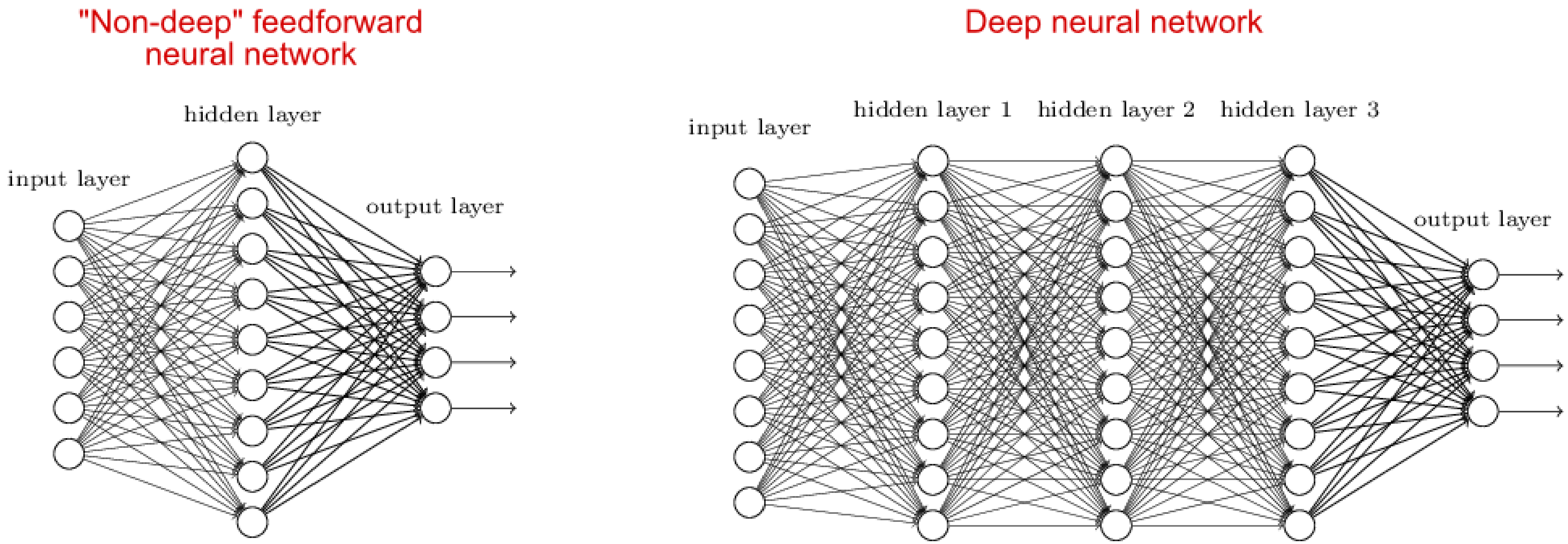
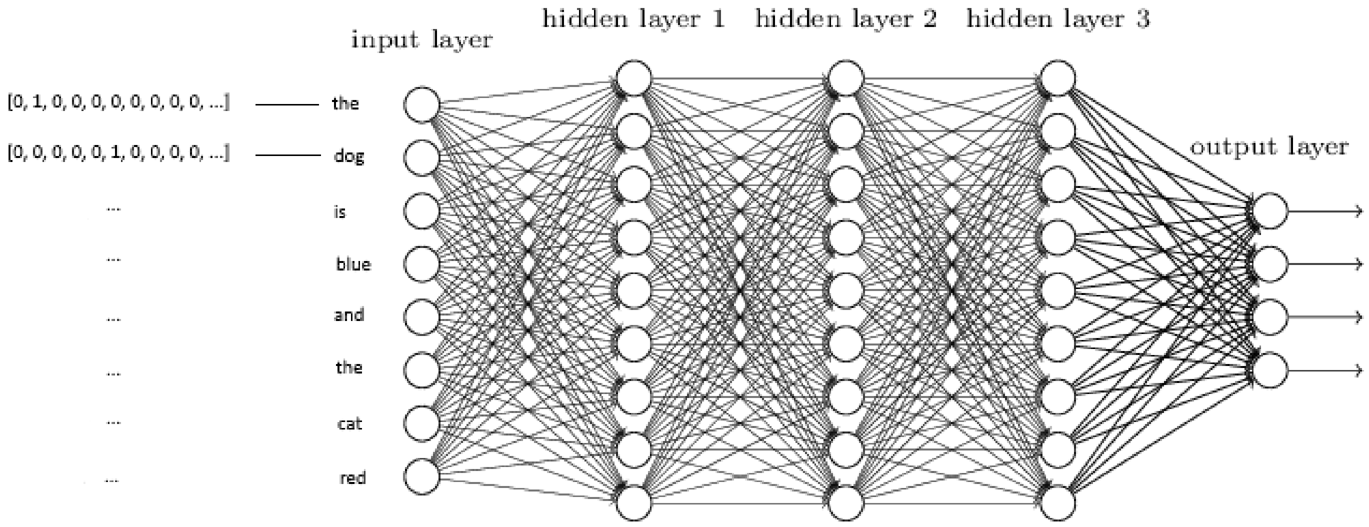


Illustration adapted from: http://neuralnetworksanddeeplearning.com/chap5.html

INPUT

- One-hot-encoding
- Vector dimension = vocabulary size
- No semantic knowledge
- Cat & dog have no relation



WORD EMBEDDINGS

"*You shall know a word by the company it keeps*" (Firth, 1957)

- Dense vector representations of words (Mikolov et al., 2013)

  [0.4232, 0.1212, 0.9483, ...]

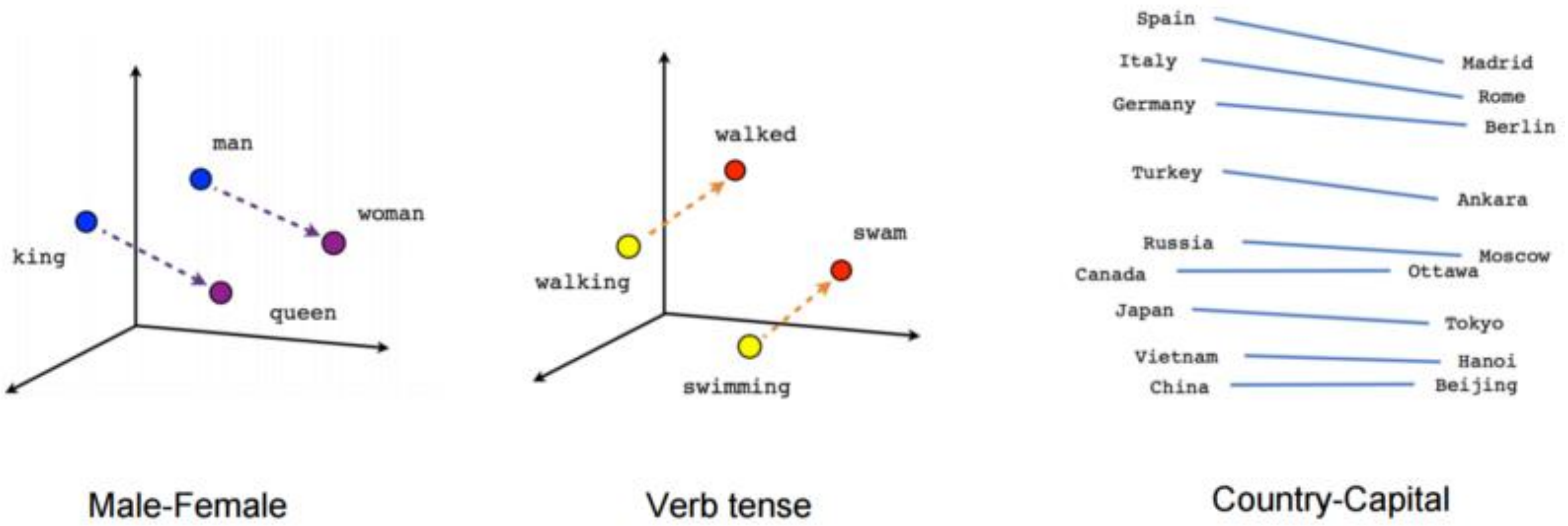


Illustration from: https://qph.ec.quoracdn.net/main-qimg-e8b83b14d7261d75754a92d0d3605e36.webp

ALTERNATIVE EMBEDDINGS

- Character embeddings (Santos and Zadrozny, 2014)
  - Sum of the embeddings of each n-gram
  - Lexical & morphological features as n-grams

< "wal", "alk", "lki", "kin", "ing", "walking">

[0.695, 0.114, ...] [0.432, 0.122, ...] [0.792, 0.094, ...] [0.932, 0.663, ...] [0.575, 0.123, ...] [0.955, 0.219, ...]

- Sentence/paragraph embeddings
  - Special combination (concatenation, addition, etc.) of word embeddings in a sentence/paragraph
  - Learned embedding representation through general language tasks

THE FEASIBILITY STUDY

REVENUE PREDICTION ON SEC FILINGS

- **Hypothesis:** it is possible given the natural language in the Management Discussion and Analysis (Item 7) of 10-K SEC filings to predict revenue increase or decrease of the following 10-K using a simple neural network architecture

- **Why?**
  - SEC filing analysis is time consuming
  - Prove that even complex language used in SEC filings can be processed by a machine
  - Potential automatic detection of outliers
  - Show future implications of what's possible

SIMPLE ARCHITECTURE

- Binary classification task
- Single LSTM architecture based on (Zaremba et al., 2015)
- Used pre-trained word embeddings
  - LexVec word embeddings (Salle et al., 2016)
  - English Wikipedia 2015 + NewsCrawl
  - 7B tokens, 368,999 words, 300 dimensions

TRAINING

- Dataset: 2903 positive & 2903 negative filings
- Train 80%, test 10%, validation 10%
- Trained on CPU (under 3 hours)
- Hyperparameters:
  - LSTM-units – depth of the network
  - Batch size – # of training examples fed into network per epoch
  - Doc length - # of words used from MDA
  - Epochs - # of traversals of training set
  - Iteration - # of times batches are fed into model for training
  - Learning rate – amount of adjustment to weights (lower = slower)
  - Learning rate decay – decrease of learning rate per epoch

RESULTS AND EVALUATION

PRELIMINARY RESULTS

Network	Units	Batch Size	Doc. length	Epochs	Iterations	Learning rate	Learning rate decay	Accuracy
LSTM	512	64	100	1	15	0.1	1/2	2.4%
LSTM	1024	64	100	1	100	0.1	1/2	79%
LSTM	256	64	1000	1	10	0.1	1/2	32.4%
LSTM	128	64	5000	5	5	0.1	1/2	n/a Computationally infeasible

EVALUATION

- Which features worked well?
  - LSTM units, document length

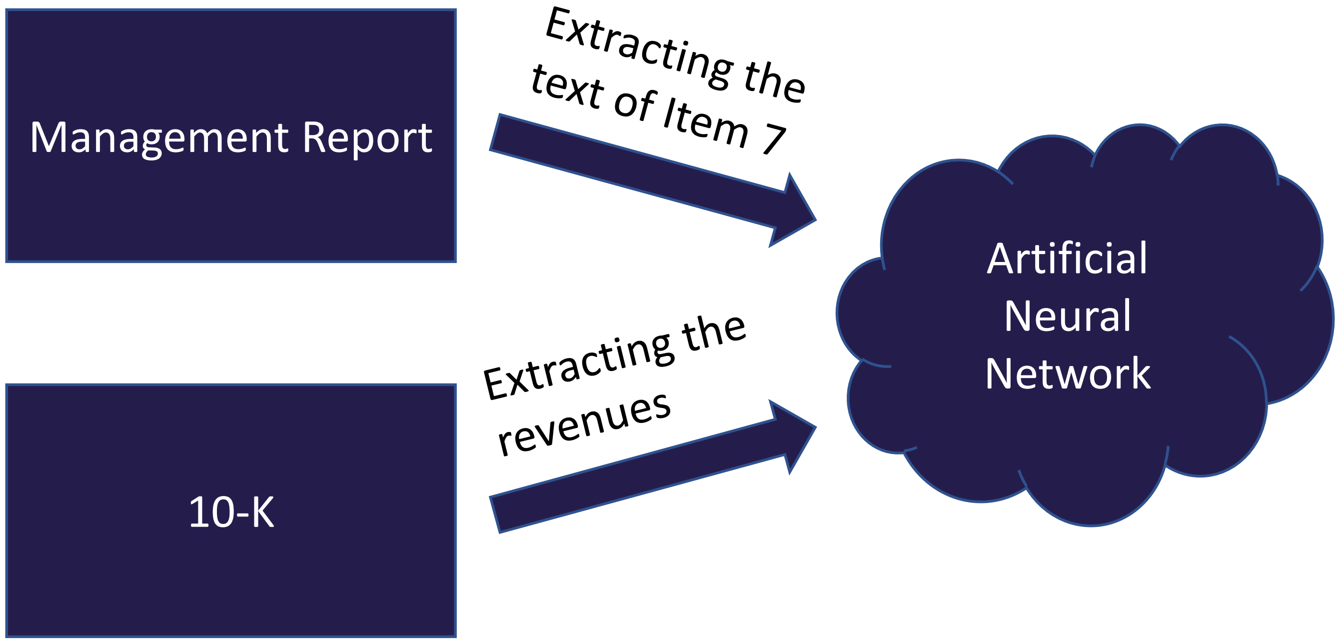
- Which didn't?
  - Batch size, epochs, iterations

EVALUATION

- What do the results tell us?
  - Inconclusive results = no baseline
  - Not enough data
  - Computationally expensive
  - Unstructured data difficult to structure
  - Preprocessing expensive (time consuming)
  - Room for improvement with…
    - structured data
    - advanced NLP approaches
    - more computational power

PROBLEMS AND OBSTACLES

DATA SOURCES FOR TRAINING THE AI



MANAGEMENT REPORT (MDA)

- ❌ Not available as XBRL
- ❌ Not available as XML
- ❌ No API
- ❌ Item 7 extraction from text
- ❌ Need to implement legacy parser

MANAGEMENT REPORT (MDA)

ITEM 7: Management's Discussion and Analysis of Financial Condition and Results of Operation

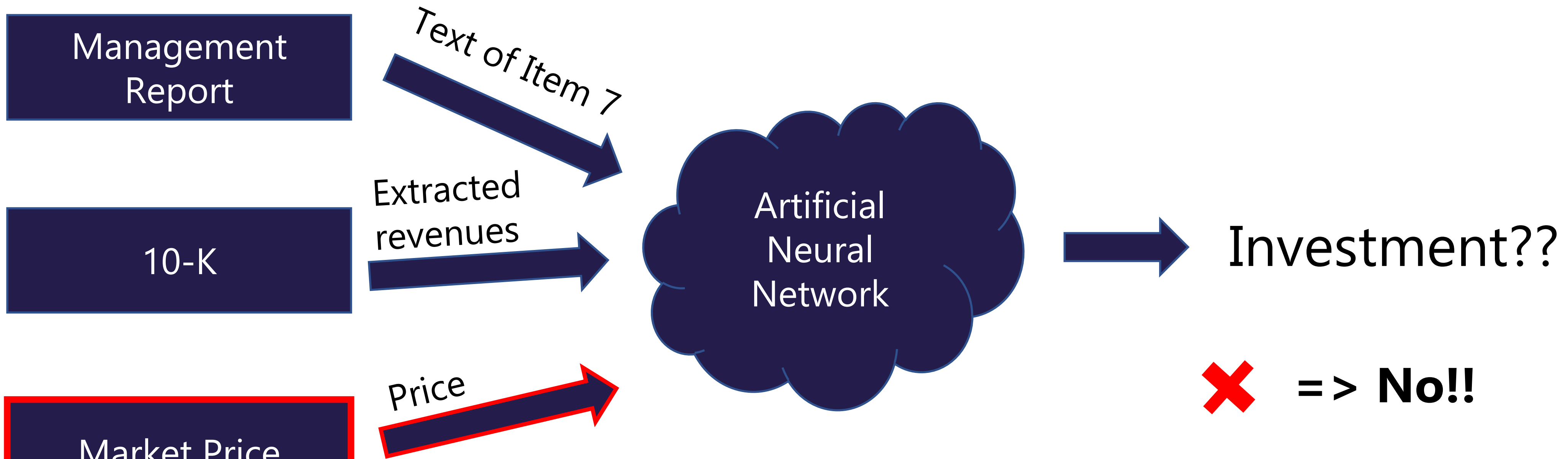
```
<p style="margin:0in 0in .0001pt;"><font size="2" face="Times New Roman" style="font-size:10.0pt;"> </font></p>
<p align="center" style="font-size:10.0pt;margin:0in 0in .0001pt;text-align:center;">22<a name="PB_22_085516_5796"></a></p>
<div style="margin:0in 0in .0001pt;"><hr size="3" width="100%" noshade align="left" style="color:#010101;"></div>
</div>
<!-- SEQ.=1,FOLIO='22',FILE='C:\JMS\105933\17-15494-1\task8552844\15494-1-bm.htm',USER='105933',CD='Aug 24 18:47 2017' -->

<br clear="all" style="page-break-before:always;">
<div style="font-family:Times New Roman;">
<p style="margin:0in 0in .0001pt;"><font size="2" face="Times New Roman" style="font-size:10.0pt;"><a href="#TableOfContents" title="Click to go to Table of Contents">Table of Contents</a>
</font></p>
<p style="margin:0in 0in .0001pt;"><font size="2" face="Times New Roman" style="font-size:10.0pt;"> </font></p>
<p style="margin:0in 0in .0001pt;"><b><font size="2" face="Times New Roman" style="font-size:10.0pt;font-weight:bold;">Item 7.  <i>Management&#146;s Discussion and Analysis of
Financial Condition and Results of Operations</i>.</font></b><a name="Item7_ManagementsDiscussionandAn_091801"></a></p>
<p style="margin:0in 0in .0001pt;"><font size="2" face="Times New Roman" style="font-size:10.0pt;"> </font></p>
<p style="margin:0in 0in .0001pt;"><b><u><font size="2" face="Times New Roman" style="font-size:10.0pt;font-weight:bold;">RESULTS OF OPERATIONS</font></u></b></p>
<p style="margin:0in 0in .0001pt;"><font size="2" face="Times New Roman" style="font-size:10.0pt;">We manufacture, market and sell beauty products including those in the skin care, makeup,
fragrance and hair care categories which are distributed in over 150 countries and territories.  <font color="black" style="color:black;">The following table is a comparative summary
of operating results for fiscal 2017, 2016 and 2015 and reflects the basis of presentation described in </font><i>Item 8. Financial Statements and Supplementary Data &#150; Note 2 &#150;
Summary of Significant Accounting Policies </i>and<i> Note 21 &#150; Segment Data </i>and<i> Related Information</i><font color="black" style="color:black;"> for all periods
presented.  Products and services that do not meet our definition of skin care, makeup, fragrance and hair care have been included in the &#147;other&#148; category.</font></font></p>
<p style="margin:0in 0in .0001pt;"><font size="2" face="Times New Roman" style="font-size:10.0pt;"> </font></p>
<table border="0" cellspacing="0" cellpadding="0" width="100%" style="border-collapse:collapse;">
<tr>
<td width="55%" valign="bottom" style="padding:0in 0in 0in 0in;width:55.5%;">
<p style="margin:0in 0in .0001pt;"><b><font size="1" face="Times New Roman" style="font-size:1.0pt;font-weight:bold;"> </font></b></p>    </td>
<td width="2%" valign="bottom" style="padding:0in 0in 0in 0in;width:2.5%;">
<p align="center" style="margin:0in 0in .0001pt;text-align:center;"><b><font size="1" face="Times New Roman" style="font-size:1.0pt;font-weight:bold;"> </font></b></p>    </td>
<td width="41%" colspan="8" valign="bottom" style="border:none;border-bottom:solid windowtext 1.0pt;padding:0in 0in 0in 0in;width:41.0%;">
<p align="center" style="margin:0in 0in .0001pt;text-align:center;"><b><font size="1" face="Times New Roman" style="font-size:8.0pt;font-
weight:bold;">Year Ended June 30</font></b></p>    </td>
<td width="1%" valign="bottom" style="padding:0in 0in 0in 0in;width:1.0%;">
<p align="center" style="margin:0in 0in .0001pt;text-align:center;"><b><font size="1" face="Times New Roman" style="font-size:1.0pt;font-weight:bold;"> </font></b></p>    </td>    </tr>
<tr>
<td width="55%" valign="bottom" style="padding:0in 0in 0in 0in;width:55.5%;">
<p style="margin:0in 0in .0001pt;"><b><font size="1" face="Times New Roman" style="font-size:1.0pt;font-weight:bold;"> </font></b></p>    </td>
<td width="2%" valign="bottom" style="padding:0in 0in 0in 0in;width:2.5%;">
```

where is the end ??

REVENUES

✖ Revenues? = us-gaap:Revenues or
us-gaap:SalesRevenueNet or
us-gaap:SalesRevenueServicesNet or
us-gaap:SalesRevenueGoodsNet or
us-gaap:OilAndGasRevenue or
… 40 more options

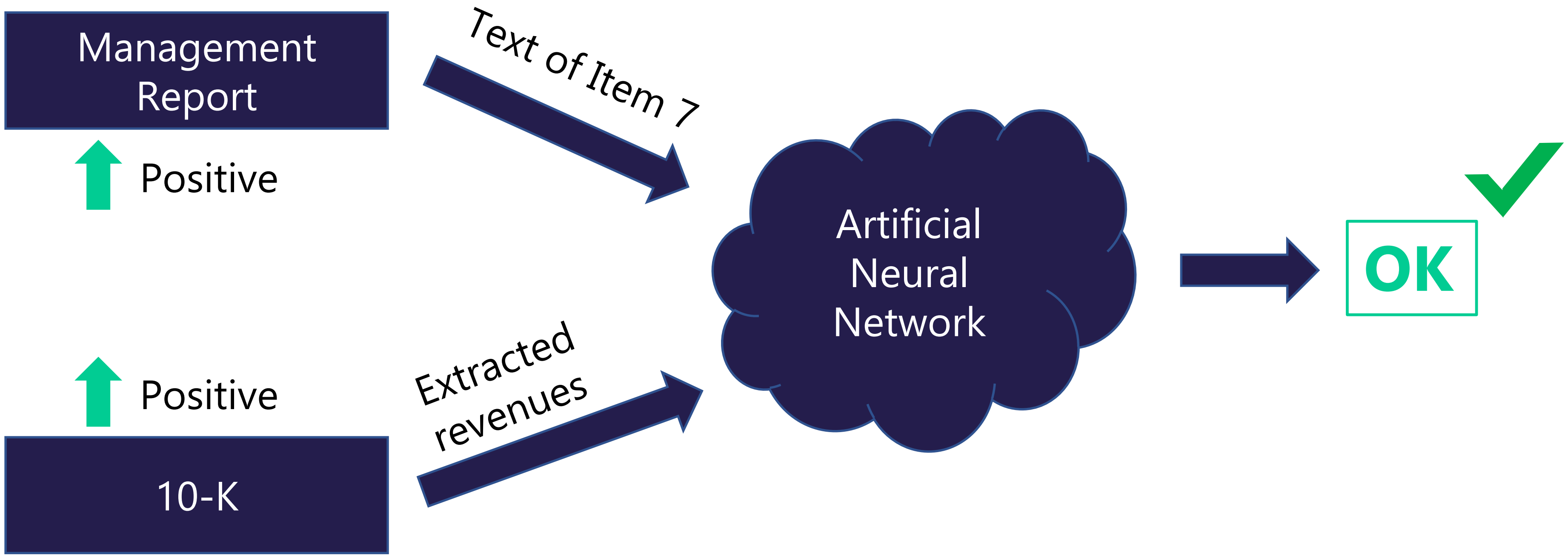
INVESTMENT DECISION?



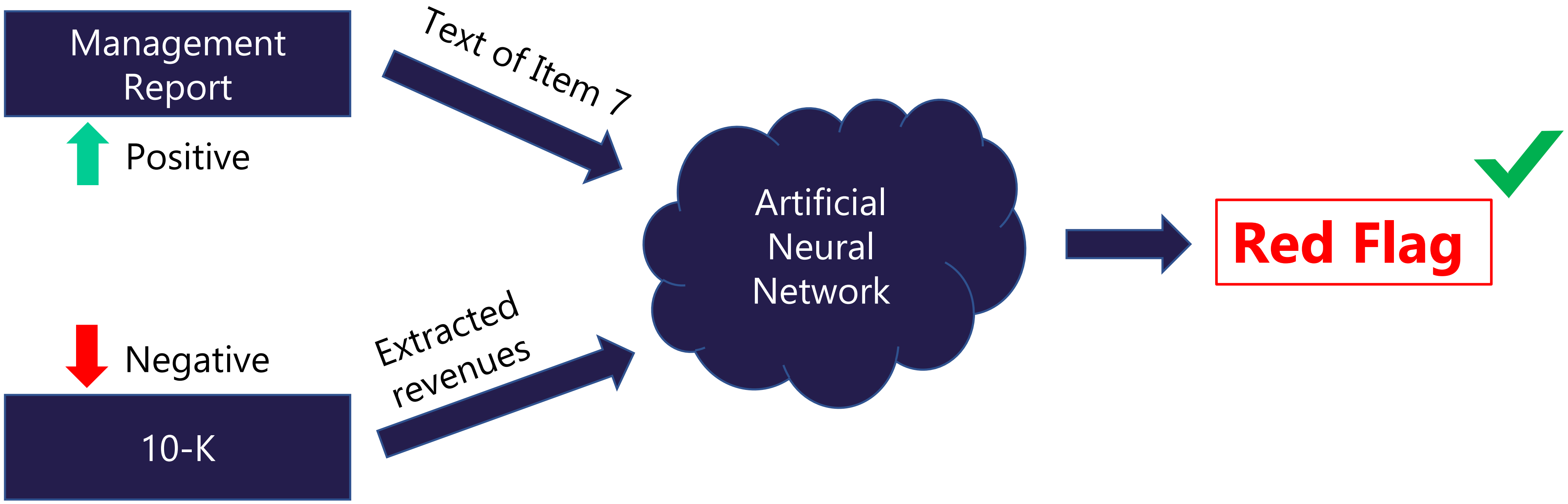
Correlation to market price missing

"Price is what you pay. Value is what you get." (Warren Buffett)

REGULATOR: OUTLIER DETECTION?



REGULATOR: OUTLIER DETECTION?



OUTLIER DETECTION

- ✗ No timely monitoring
- ✗ Need for more data
- ✓ Long-term tracking of credibility
- ✓ <u>Contributes</u> to a sound outlier detection

OUTLOOK

FUTURE APPLICATIONS

- Outlier detection; gets better with…
  - more data
  - multi-class classification
  - additional natural language sources
- Self monitoring / benchmarking
- Auditing
- Legal act reporting implications

ALTERNATIVE APPROACHES

- Instead of (manually) engineering natural language data, use data with naturally occurring markers or features

- Pros:
  - Abundance of data
  - No annotation needed (inexpensive)

- Cons:
  - Difficult to find or hypothesize such features

POSSIBLE FUTURE WORK

- **InferSent** (Conneau et al., 2017): generic sentence representations that can outperform task specific implementations
- Task: given 2 sentences, determine their relationship between [contradiction, neutral or entailment]
- Diverse semantic knowledge included in sentence representations

"***A man inspects the uniform of a figure in some East Asian country.***"

- contradiction -

"***The man is sleeping.***"

POSSIBLE FUTURE WORK

- **DisSent** (Nie et al., 2018): naturally annotated sentence relationships allow robust sentence representations; state-of-the-art performance
- Naturally occurring markers used to incorporate semantic knowledge into sentence embeddings
- Task: given 2 sentences, predict which discourse marker (and, but, because, etc.) was used

"***She's late to class ___ she missed the bus.***"

"***She's good at soccer ___ she missed the goal.***"

POSSIBLE FUTURE WORK

- Train task-specific word/sentence embeddings
- More complex architectures:
  - Attention Networks (Yang et al., 2016), which focus on "important" parts of a sentence
  - Gated Recurrent Unit (Cho et al., 2014), which is a simpler variant of LSTM
- Train on GPU
- More/better hyperparameter tuning
- Quickly advancing field with many new approaches possible

THANK YOU