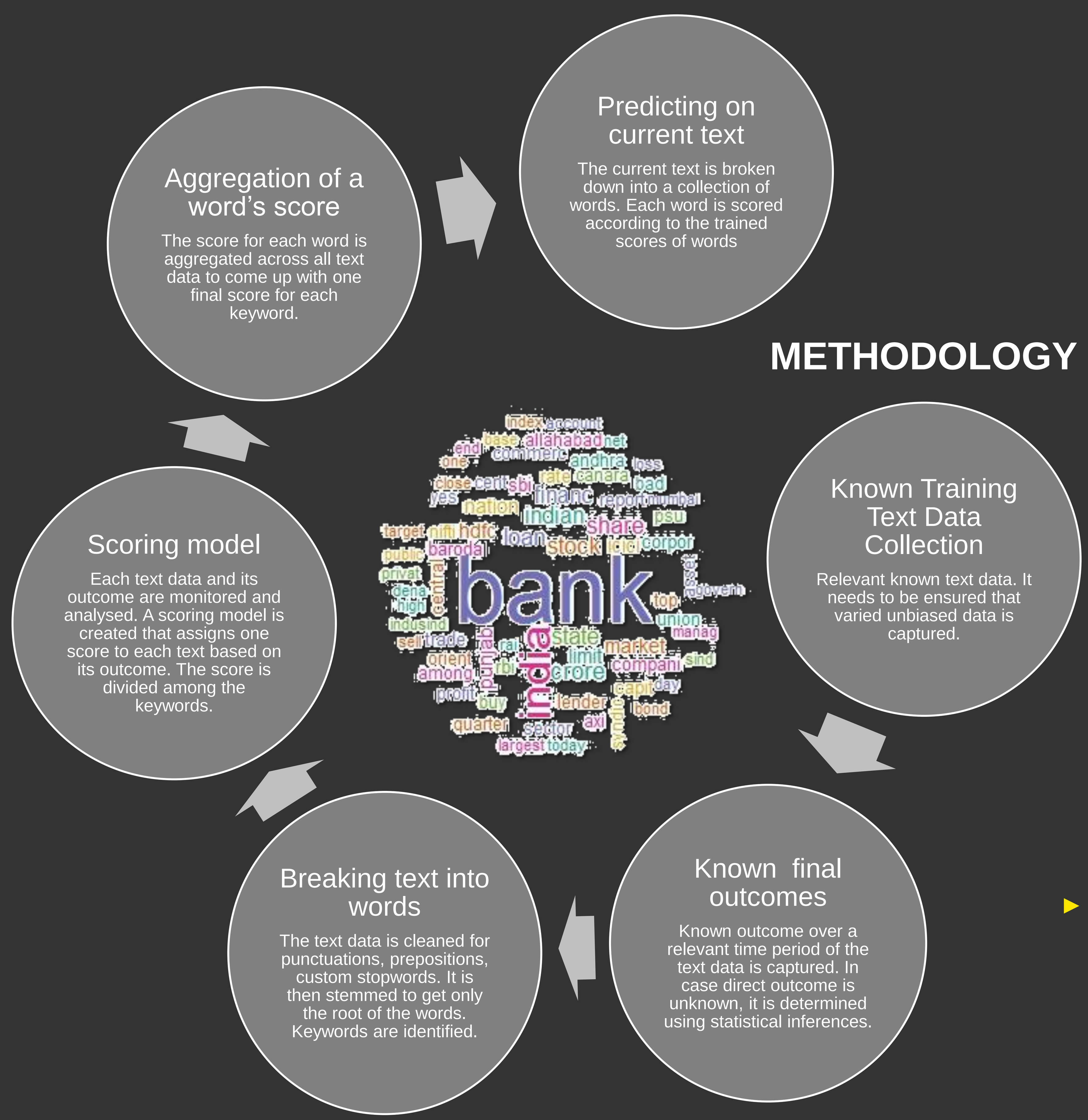
Prediction using Sentiment & Text Analytics

Prabakar Rajasekaran, Tvisha Gupta, Manoj Kumar Reddy Gadipally
GDS

"When words are scarce, they are seldom spent in vain" – William Shakespeare

The answer to "What is the likely outcome of a given scenario?" depends on an analysis of multiple training scenarios. It involves the selection of appropriate predictors and creating a model to provide an estimate of the probable outcome. With well defined and structured variables, this task is fairly existent in practice. But what if the scenarios we have are not in any structure but rather in the form of text data? The first step here – interpret the text! With the industry dealing with voluminous amount of text data in the form of reports, news, forms, views, it is essential to have a methodology to interpret the data as much as possible avoiding any loss of information. At the same time it is important that it is done by means of an automated engine so that the process is intuitive, fast, consistent and is a scientific and systematic way rather than human manual judgement. This ensures a higher quality of interpretation at a lower cost with reduced time. This brings the need for text and sentiment analytics.

METHODOLOGY



MODEL DOCUMENT INFERENCE

- A modelling exercise is incomplete without a thorough documentation process which often takes the majority of the time. Based on a lot of different premise in the form of statistical tests and presence of certain qualitative checklists, it is decided whether the model is approved or rejected.
- A dictionary of positive and negative elements in a document is created.
- Words in each premise determine its sentiment. Based on the sentiments associated with the premise, a score is given to it. This score is divided amongst the words present in the premise. This helps us in preparing the training model.
- With this training model, each keyword in a premise and test in a model document is assigned a score. Scores for all the words are added to determine the overall sentiment associated with premise. This is done for every premise.
- A frequency table is created to see the number of positive and negative premise in the document.
- Based on whichever is higher – positive or negative premise, it is decided whether the model should be approved or rejected.

STOCK MOVEMENT PREDICTION

- Financial markets are primarily driven by investor sentiments and how the investors perceive a news item to be positive or to be a negative one.
- A database of historical news items is created by scraping web pages. At the same a database of half hourly stock price data is created. Each news item is mapped to the next 10 half hourly returns that would have been witnessed after the announcement of the news.
- A score is assigned to each item by the formula

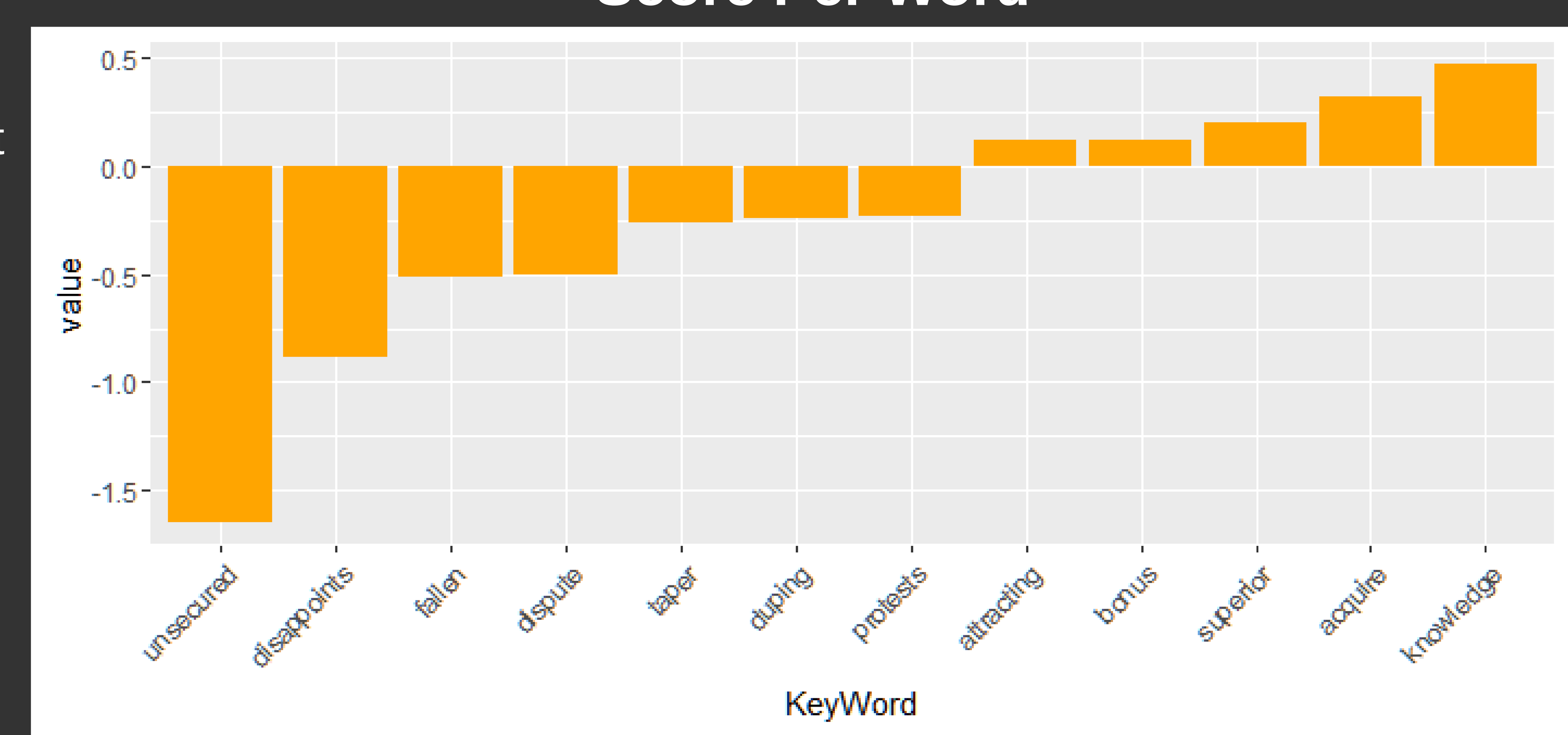
  Maximum Return * Maximum Positive Streak + Minimum Return * Maximum Negative Streak
- The score is divided into the keywords. Scores from multiple news items give an aggregated score per keyword.
- A stock is selected whose movement needs to be predicted. Relevant recent news for the stock are searched online.
- Keywords present in the news item are assigned the score from the training model. The aggregate score determines the sentiment of the news item and hence the likely movement that the stock price would have.

OBSERVATIONS

- Negative news is quicker to show its effects on the stock prices than a positive news – a possible reason for this is that investors get worried about possible losses and sell off their stocks very quickly. But to invest more, they take time to analyse and hence the movements are not immediately reflected.
- In case of document inference, modelers never accept a premise. They just "do not have sufficient evidence to reject it". As a result, they avoid the presence of positive words. Hence just like stock price movement, the accurate detection rate of negative words is higher than that of positive words.
- Stock price movement is at a lot of times not directly correlated with the news items. This leads to situations where the movement suggested by the returns is different from that suggested by the words present. However at an aggregate level, this issue is minimized

PREDICTION ACCURACY RATIO – 53% (More than 65% for negative prediction)

Score Per Word



The better the question. The better the answer.
The better the world works.

EY
Building a better working world