

# CLASSIFICATION

CS5604 Information Storage and Retrieval - Fall 2016

Virginia Polytechnic Institute and State University

Blacksburg, Virginia 24061

Professor: E. Fox

Presenters:

Saurabh Chakravarty,


Eric Williamson

December 1, 2016



# TABLE OF CONTENTS

---

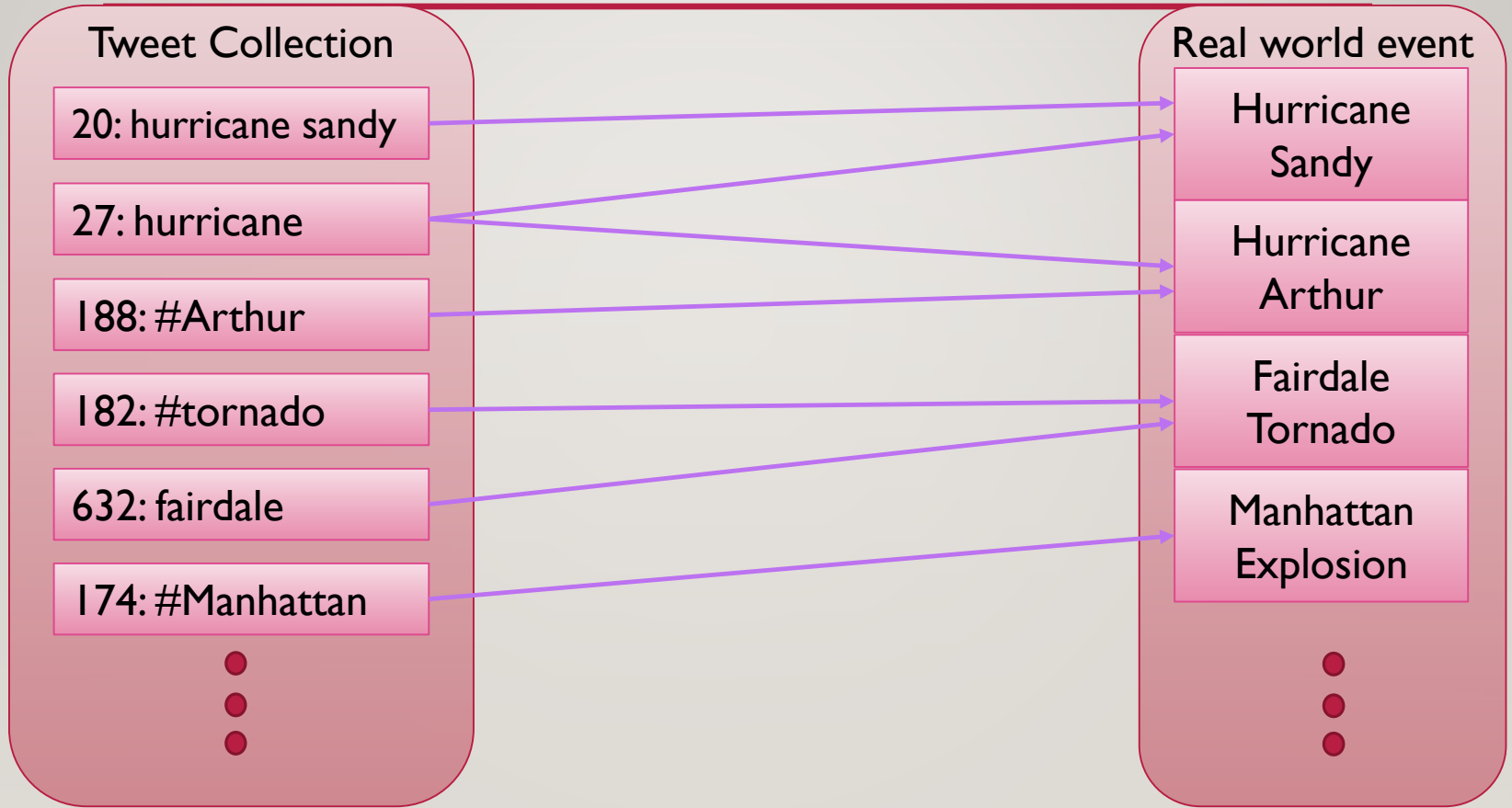
- Problem Definition
  - High-Level Architecture
  - Data-Retrieval and Processing
  - Classification
  - Experimental Results
  - Conclusion and Future work
- 

# PROBLEM STATEMENT

---

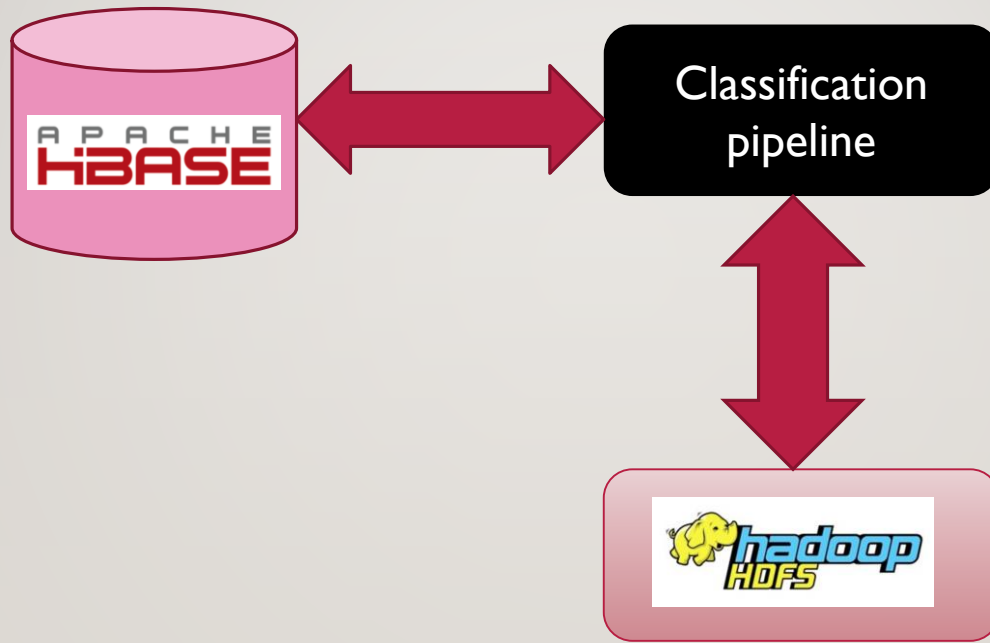
- Given the tweets in the GETAR and IDEAL collections and a set of real world events, determine which tweets belong to each real world event.

# PROBLEM STATEMENT



# HIGH LEVEL ARCHITECTURE

---



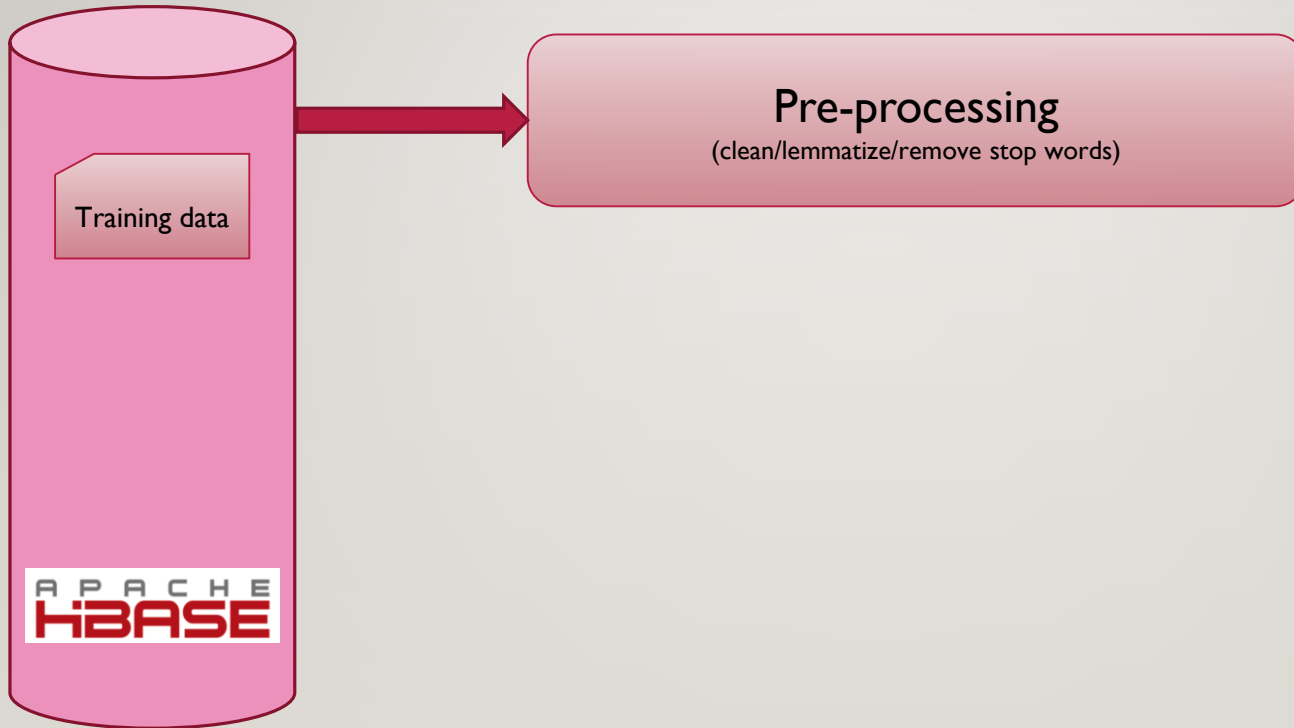
# TRAINING PIPELINE

---



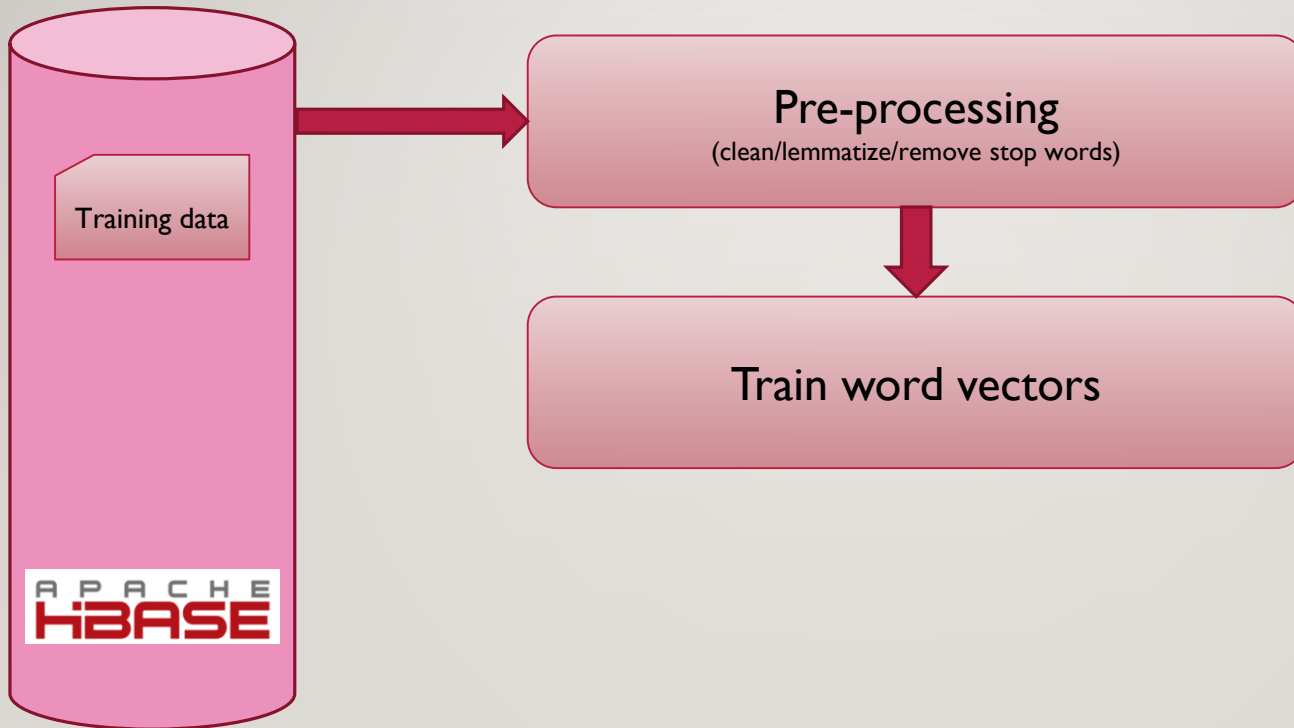
# TRAINING PIPELINE

---



# TRAINING PIPELINE

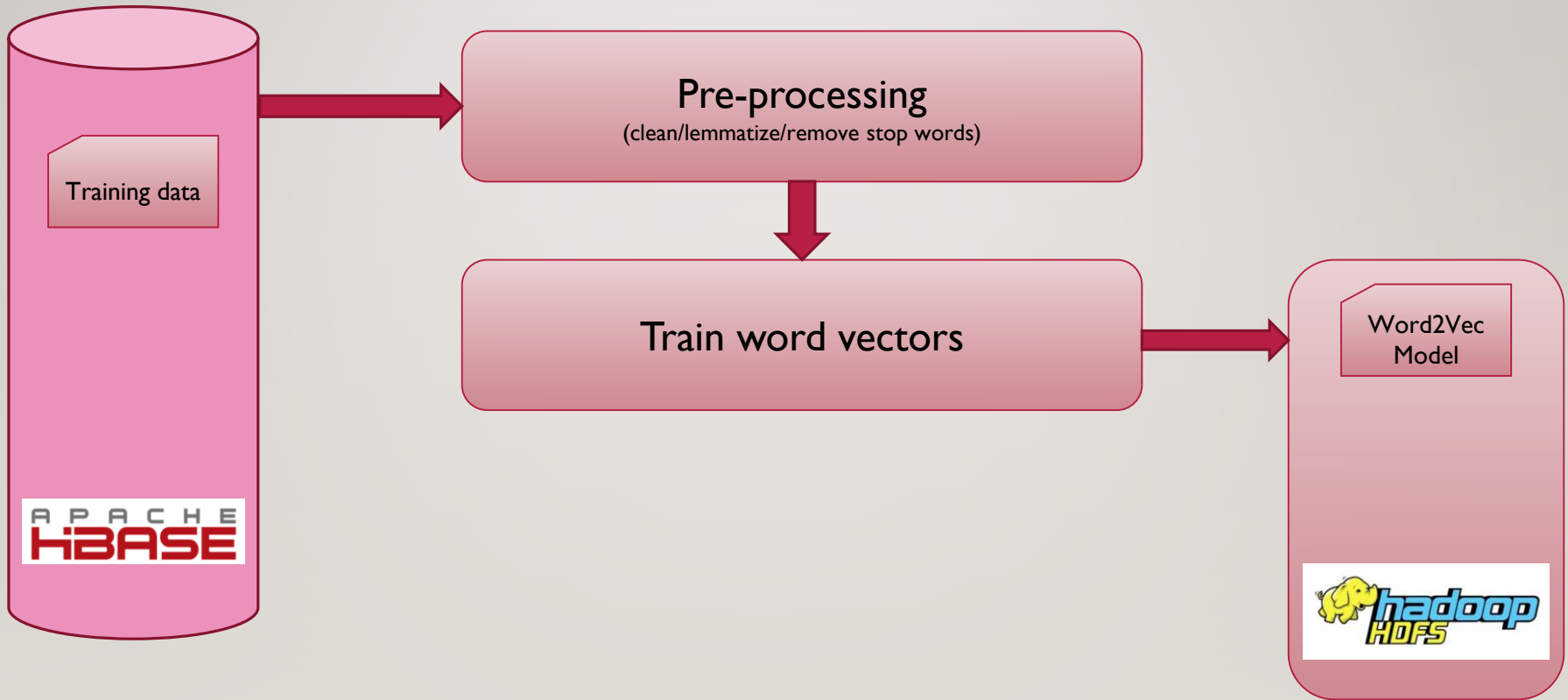
---



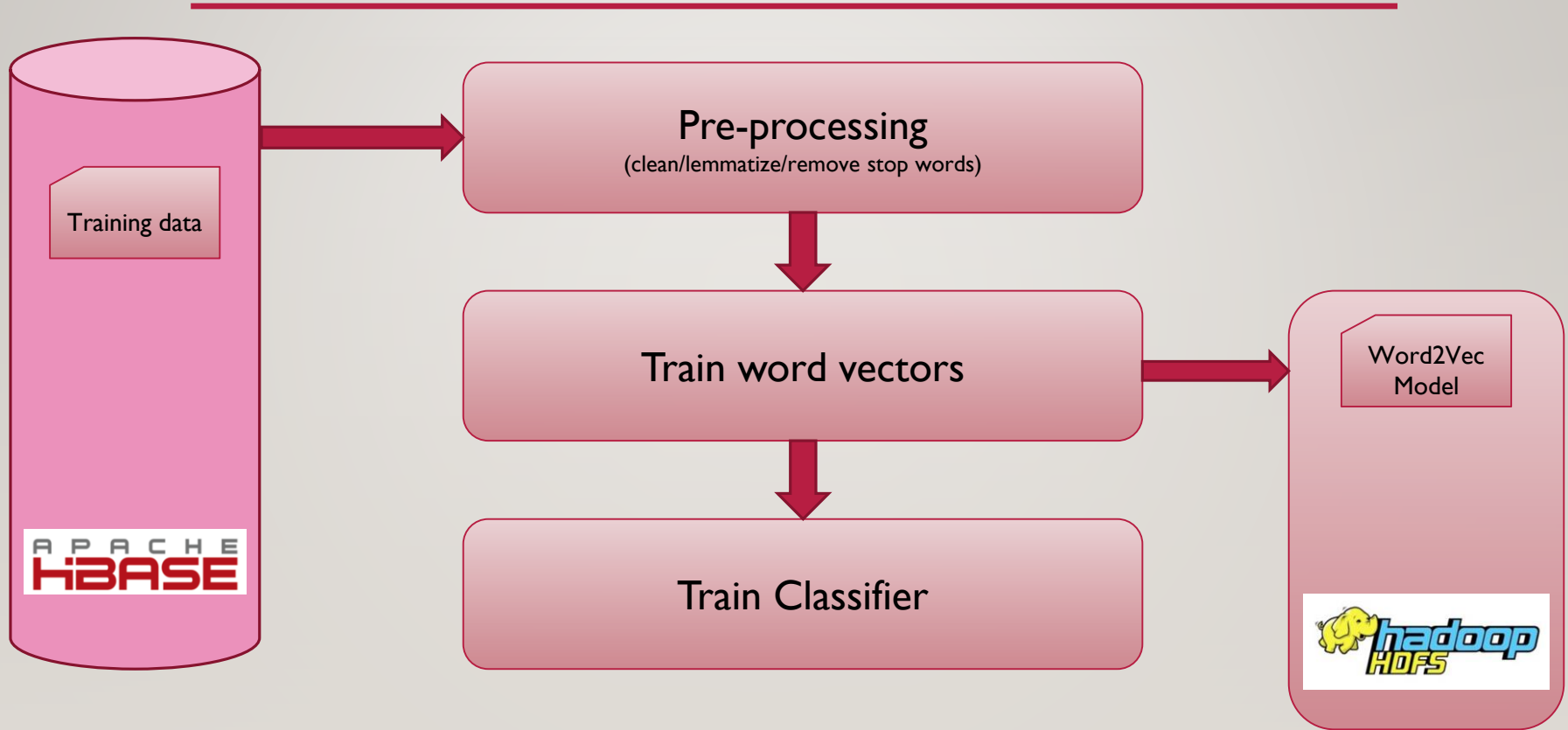


# TRAINING PIPELINE

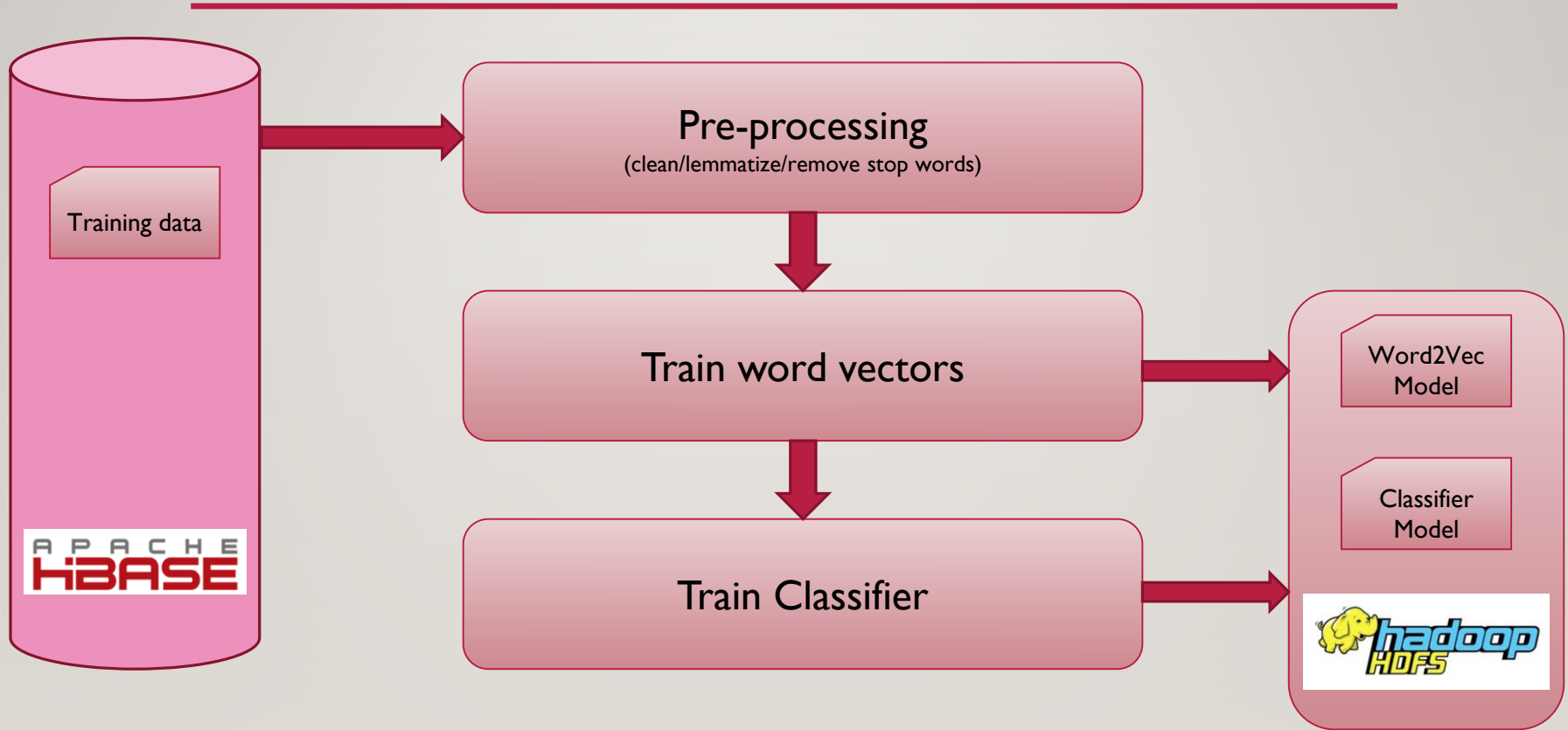
---



# TRAINING PIPELINE



# TRAINING PIPELINE

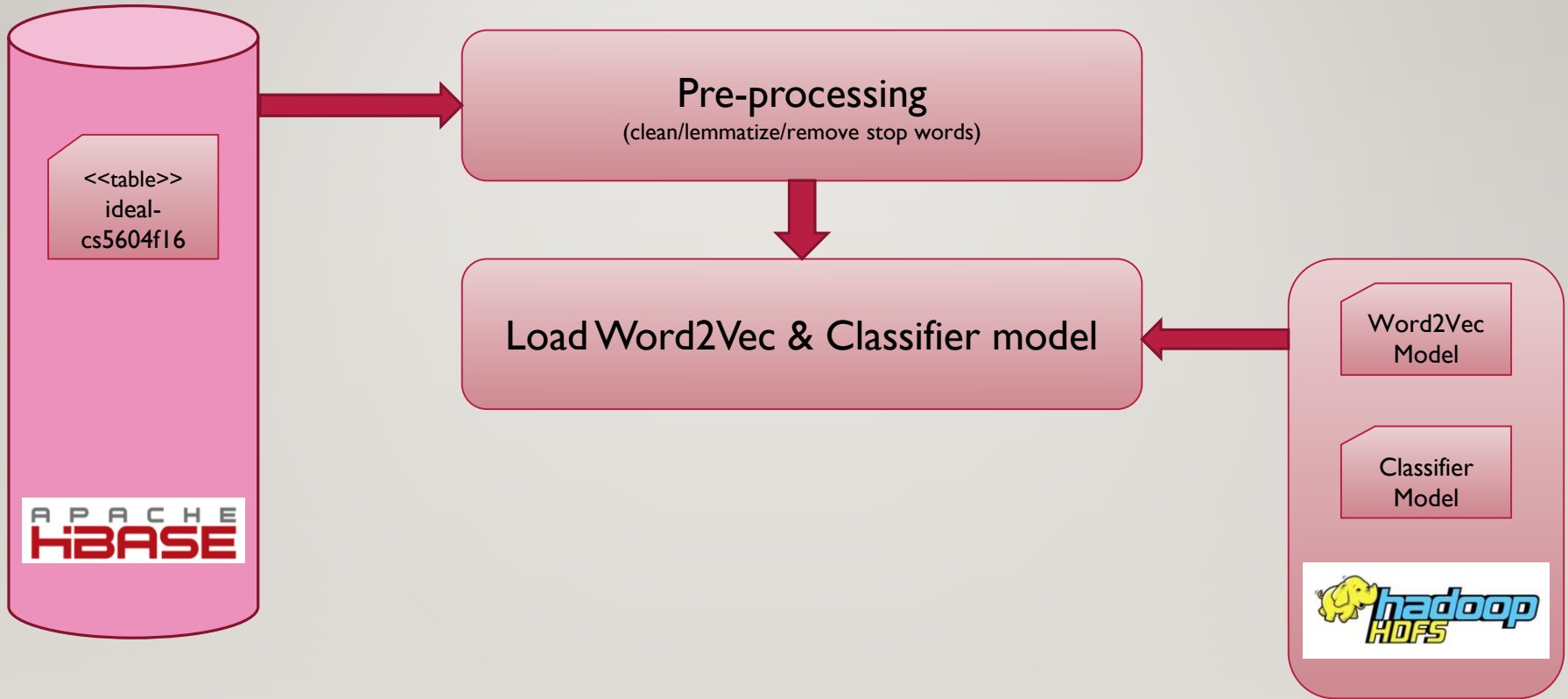


# PREDICTION PIPELINE

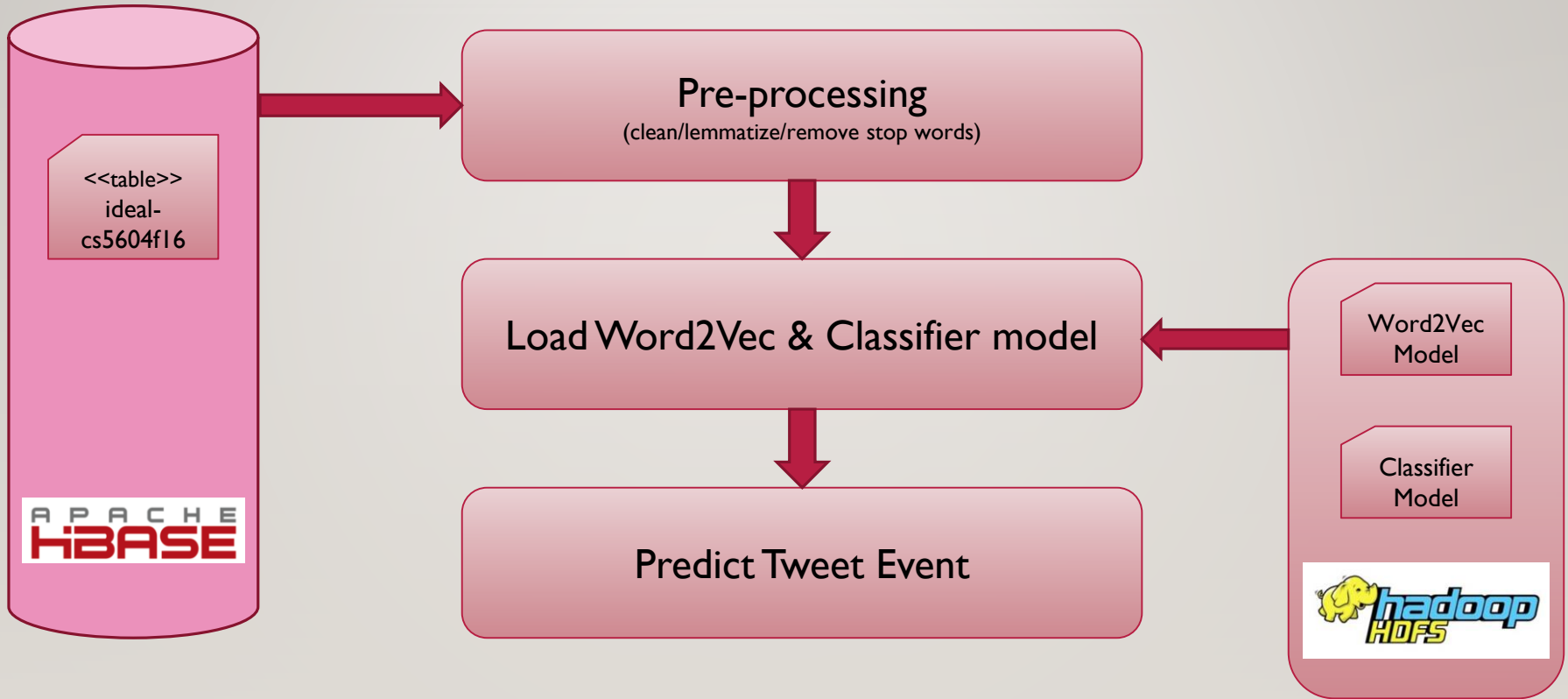
---



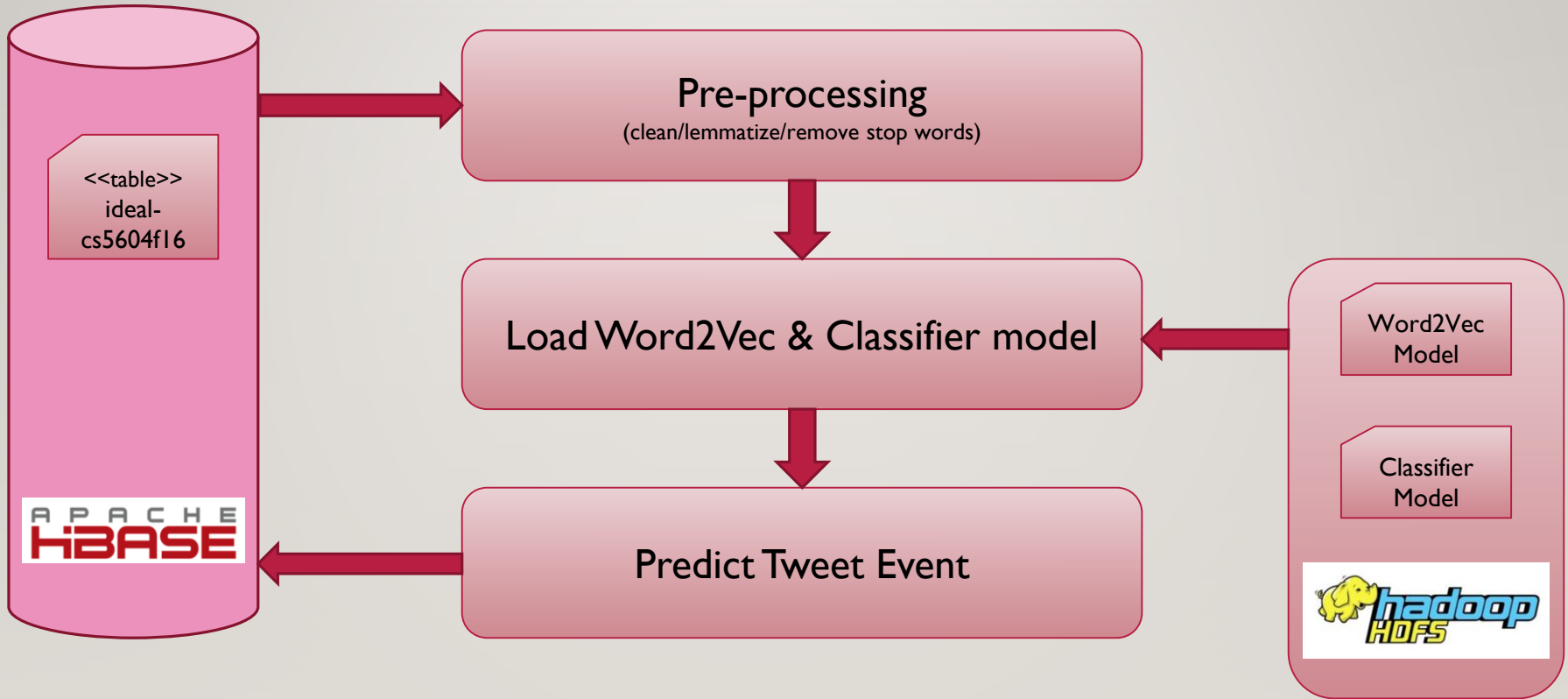
# PREDICTION PIPELINE



# PREDICTION PIPELINE



# PREDICTION PIPELINE



# DATA RETRIEVAL CHALLENGES

---

- Large amounts of data (1.5 billion tweets)
- Have to avoid serial execution
- Cannot fit into memory
- Prevent reprocessing data unnecessarily



# DATA RETRIEVAL FROM HBASE

---

Retrieval Method	Description	Smaller collection performance	Larger collection performance
Spark HadoopRDD	Load data into driver and parallelize across the cluster.	Seamless reading from HBase.	Hangs and does not complete reading on collections > one million.
Batch Processing	Load one batch at a time onto the drive and parallelize across the cluster.	Slower reading due to batch overhead.	Allows classification or arbitrarily large collections.

# CHALLENGES WITH TWEETS

---

- Abbreviations and Slang (u,omg)
- Non English URLs and characters
- Misspellings

RT: @AssociationsNow A Year After Texas Explosion Federal  
Repourt Outlines Progress on Fertilize...  
<http://t.co/8fDbMu9asU> #meetingprofs

# TEXT-PREPROCESSING

---

- Remove URLs
- Remove the # characters
- Lemmatization using StanfordNLP
- Stopword removal

# CLEANING EXAMPLE

---

## **Raw:**

RT: @AssociationsNow A Year After Texas Explosion Federal Report Outlines Progress on Fertilize...

<http://t.co/8fDbMu9asU> #meetingprofs

## **Clean:**

year texas explosion federal report outline progress fertilize  
meetingprof

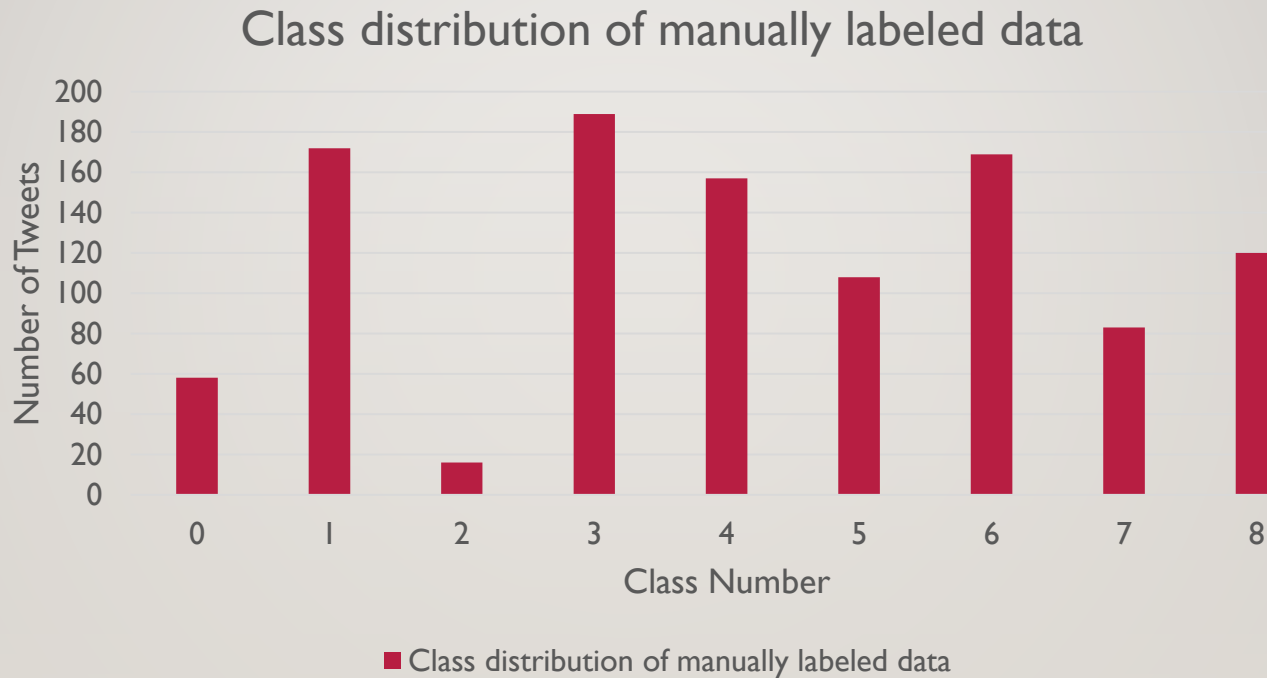
# TRAINING DATA GENERATION

---

- Random samples of collections corresponding to real world events.
- Tweet assigned a number with the class it belongs to most.
- Hand Labeled 1000 Tweets

# CLASS DISTRIBUTION FOR TRAINING AND TEST DATA

---



# TEXT CLASSIFICATION

---

- Feature selection
- Feature representation
- Choice of classifier

# A COMPARISON OF FEATURE SELECTION TECHNIQUES

Technique	Advantages	Disadvantages
Tf-idf	<ul style="list-style-type: none"><li>• Superior for small feature.</li><li>• High term removal capability.</li></ul>	Accuracy suffers for large datasets.
Mutual information	Simple to implement.	Inferior accuracy performance.
Association rules	<ul style="list-style-type: none"><li>• Fast execution.</li><li>• Very good accuracy for multi-class scenarios.</li><li>• Easy to interpret the rules</li></ul>	Prone to discovering too many rules or poorly understandable rules that hurt performance and interpretation.
Chi-square statistic	Robust accuracy and performance with large sample sets with fewer classes.	Difficulty in interpretation of when there are a large number of classes.
Within class popularity	Identifies words that are most discriminative.	Ignores the sequence of words.
Word2Vec	Captures relationships of a word with neighbors.	High computational complexity. Long training time for large sample size.

For more details, please refer the appendix section.



# COMMON FEATURE REPRESENTATION TECHNIQUES

---

- One-hot encoding
- Bag of words

## Challenges

- Large number of dimensions
- Word relationships with neighbors are not captured

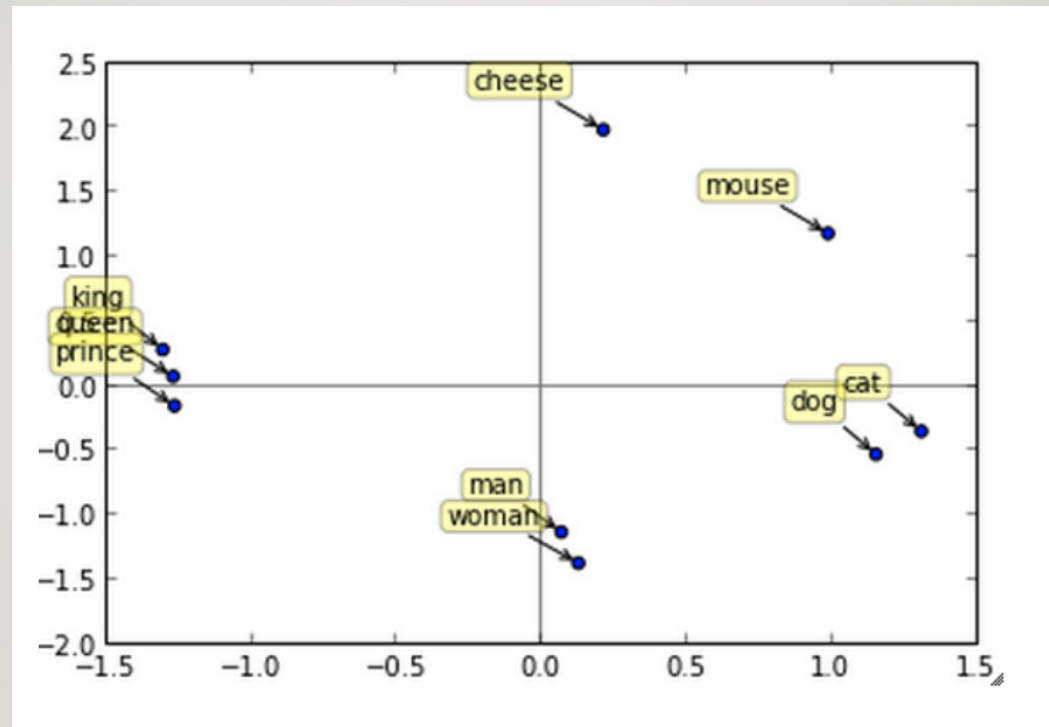
# WORD2VEC

---

- A feature selection technique.
- Captures the semantic context of a word's relation with neighbors.

For more details, refer the appendix.

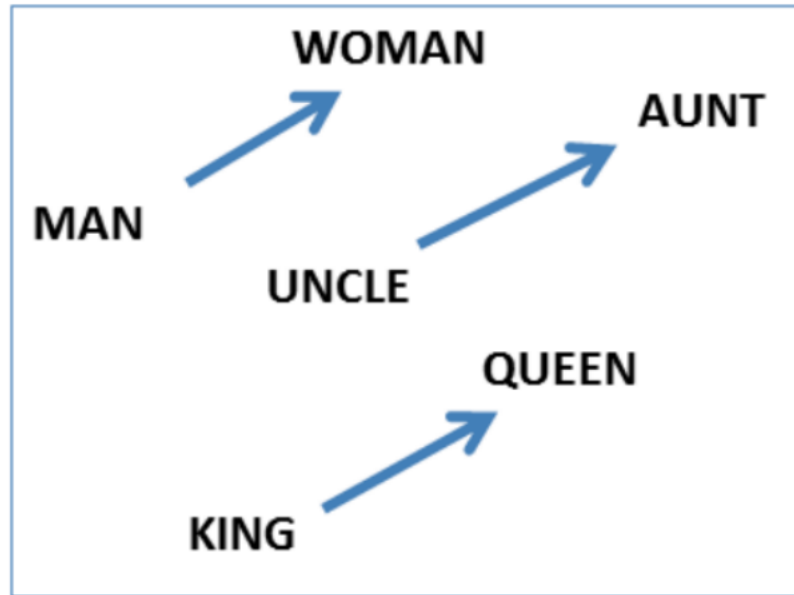
# WORD2VEC



Similar words are grouped together and closer to one another.

# WORD2VEC

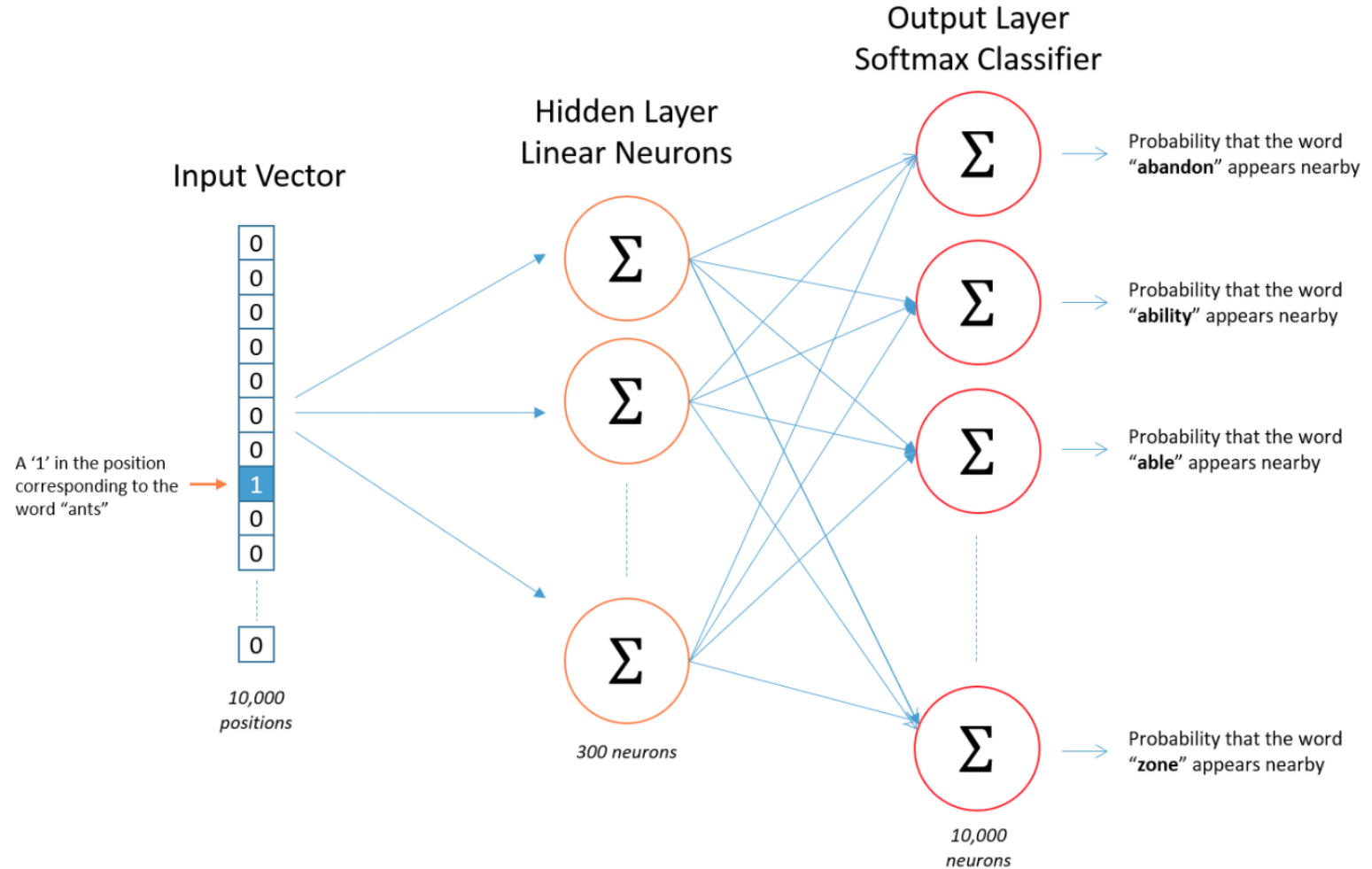
---



Word displacements are relationships between the words.

Source: *Linguistic Regularities in Continuous Space Word Representations*, Mikolov et al, 2013

# WORD2VEC



# WORD2VEC

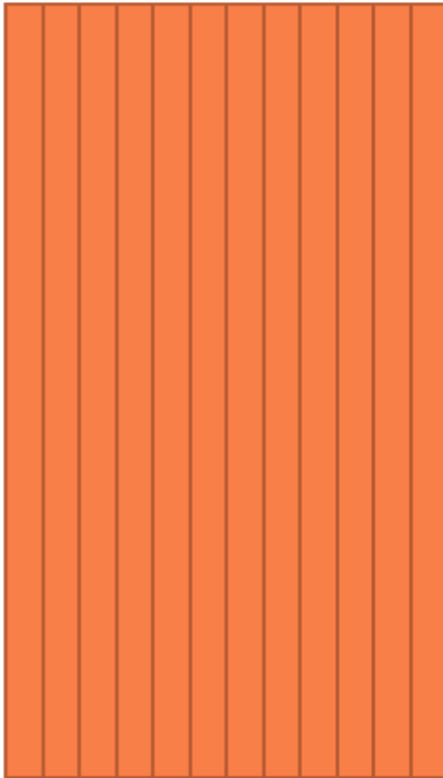
Hidden Layer  
Weight Matrix



*Word Vector  
Lookup Table!*

*300 neurons*

*10,000 words*



*300 features*

*10,000 words*



# CLASSIFIER – IMPLEMENTATION DETAILS

---

- A word feature is an average of the word vectors generated by the Word2Vec model.
- We used a vector representation with a default of 100 values.
- We chose the **multi-class logistic regression** in the Spark framework to perform classification.
- The classifier labels the **predicted class** along with the **normalized probabilities** of the other classes.

# EXPERIMENTS

---

- Effect of preprocessing
- Accuracy performance
- Runtime performance
- Probability distribution
- Class assignment



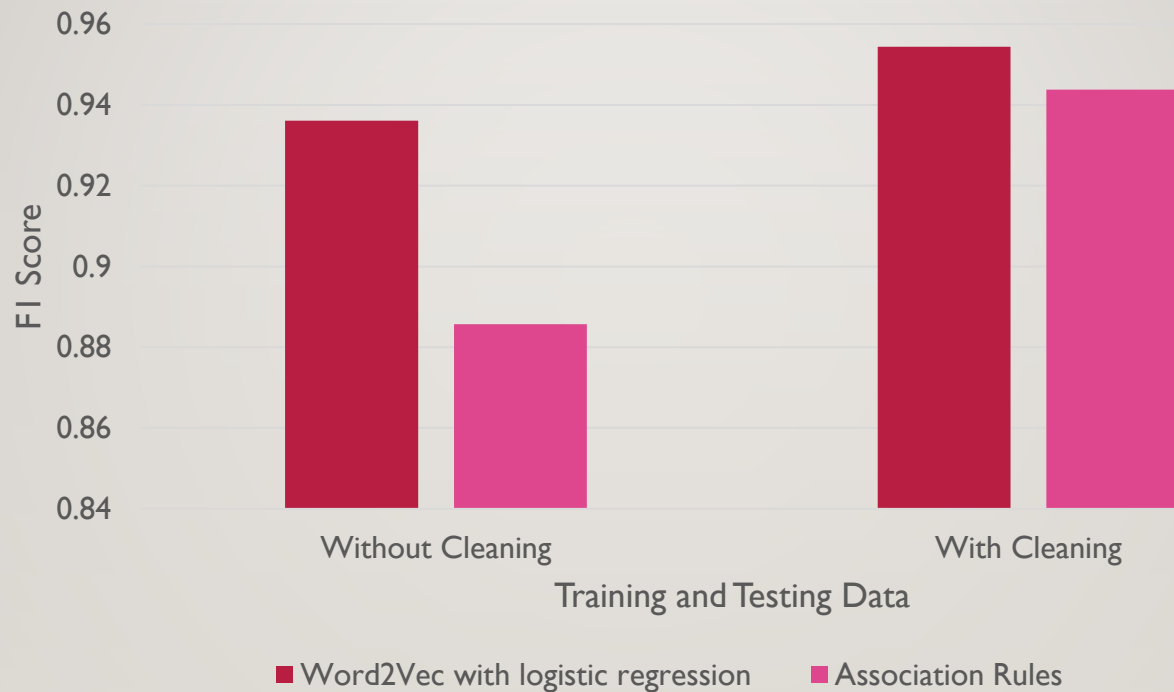
# CLEANING EXPERIMENT

---

- Determine how cleaning the data influences accuracy
- Cleaning:
  - Lemmatization
  - Stopword removal
  - Hashtag removal
- Experimental setup:
  - Split hand-labeled data:
    - 70% train
    - 30% test

# CLEANING IMPROVES ACCURACY!

---



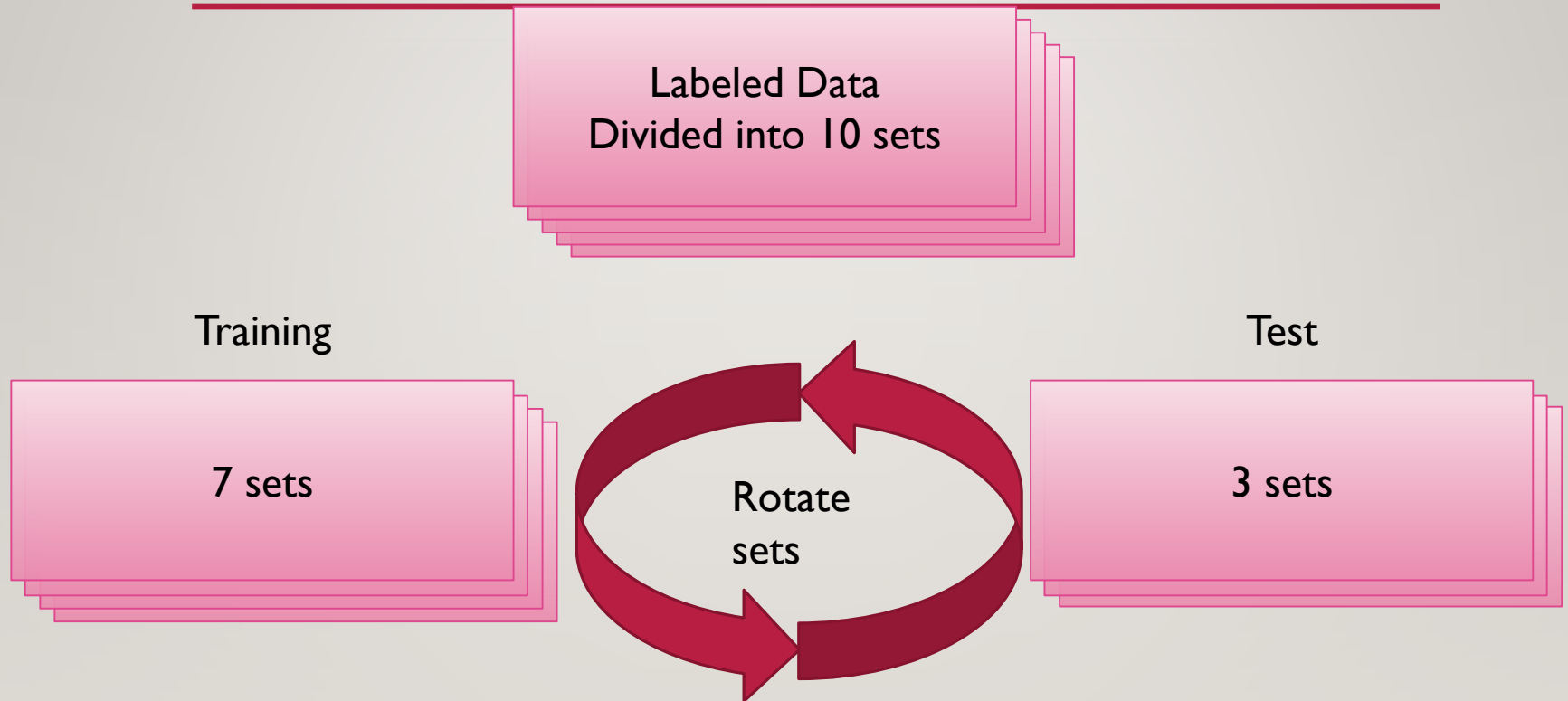
Word2Vec with logistic regression: 29% fewer misclassifications  
Association Rules: 51% fewer misclassifications

# ACCURACY EXPERIMENT

---

- Determine which classifier gives better results on labeled data
- Experimental setup:
  - Generate 10 different breakups of the labeled data
  - Calculate metrics for each classifier on same breakup
  - 9 different classes

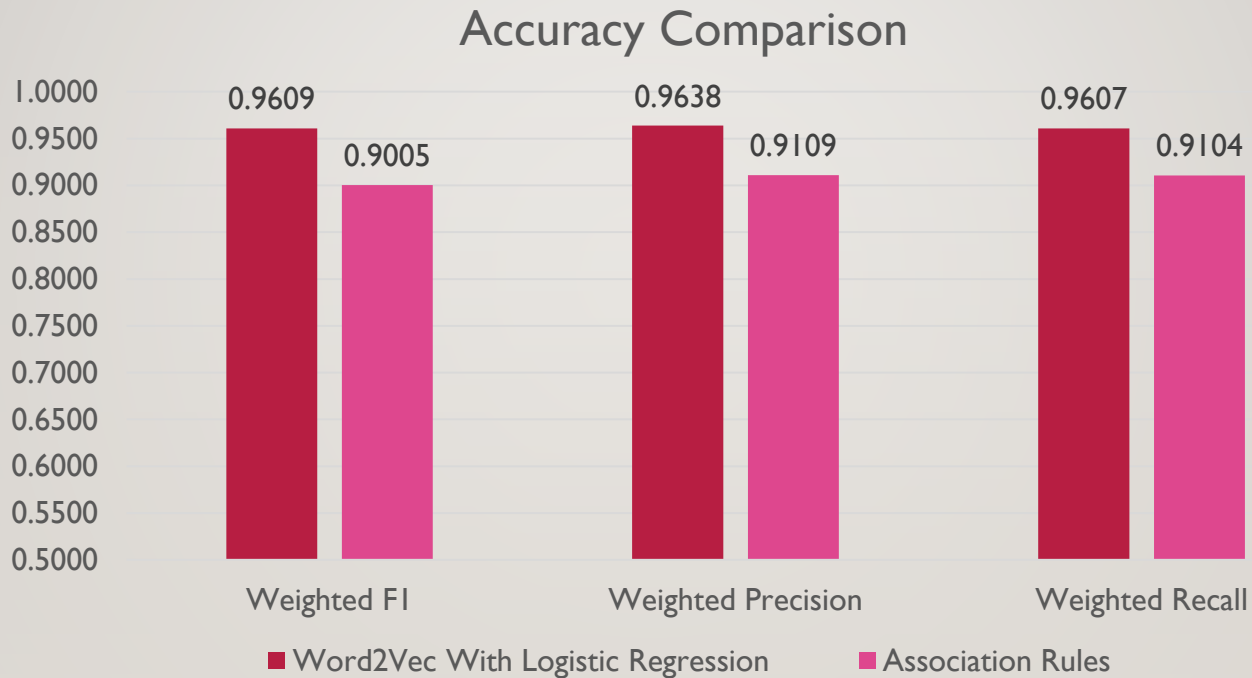
# ACCURACY EXPERIMENTAL SETUP



Generate 10 different training and test sets

# WORD2VEC OUTPERFORMS ASSOCIATION RULES

---



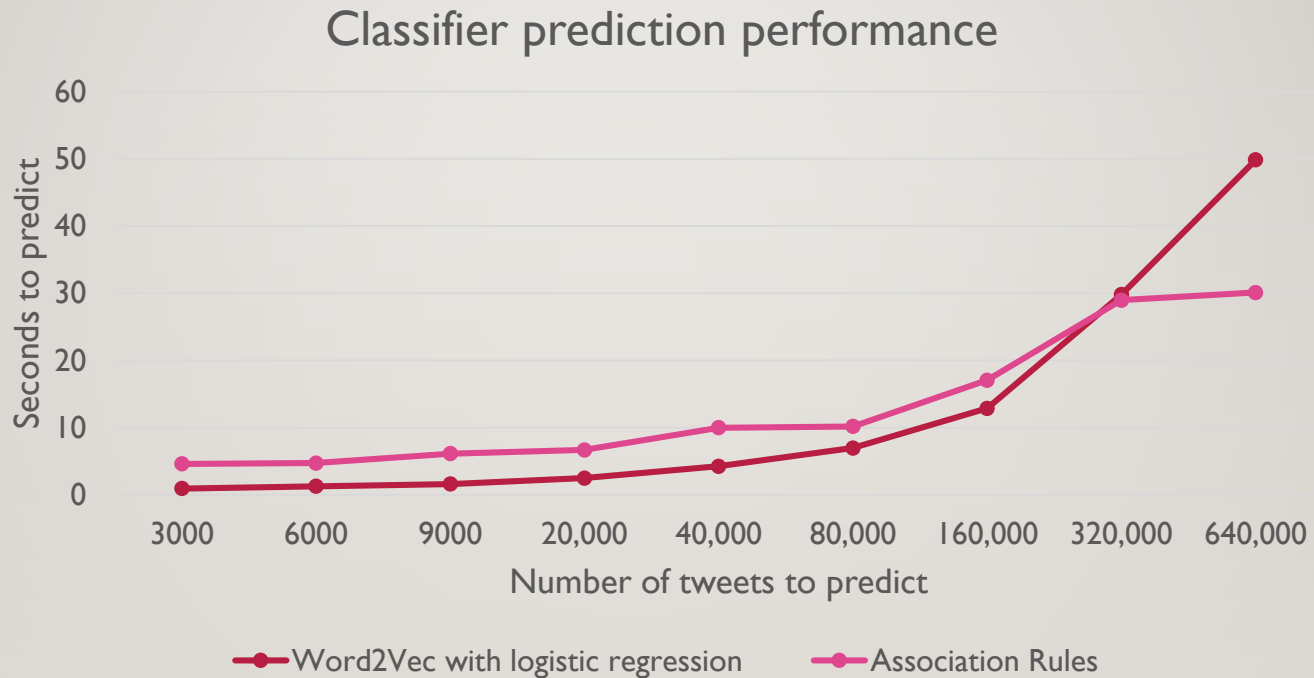
Word2Vec with logistic regression had a **6.7%** increase in FI score over association rules.

# CLASSIFIER RUNTIME PERFORMANCE

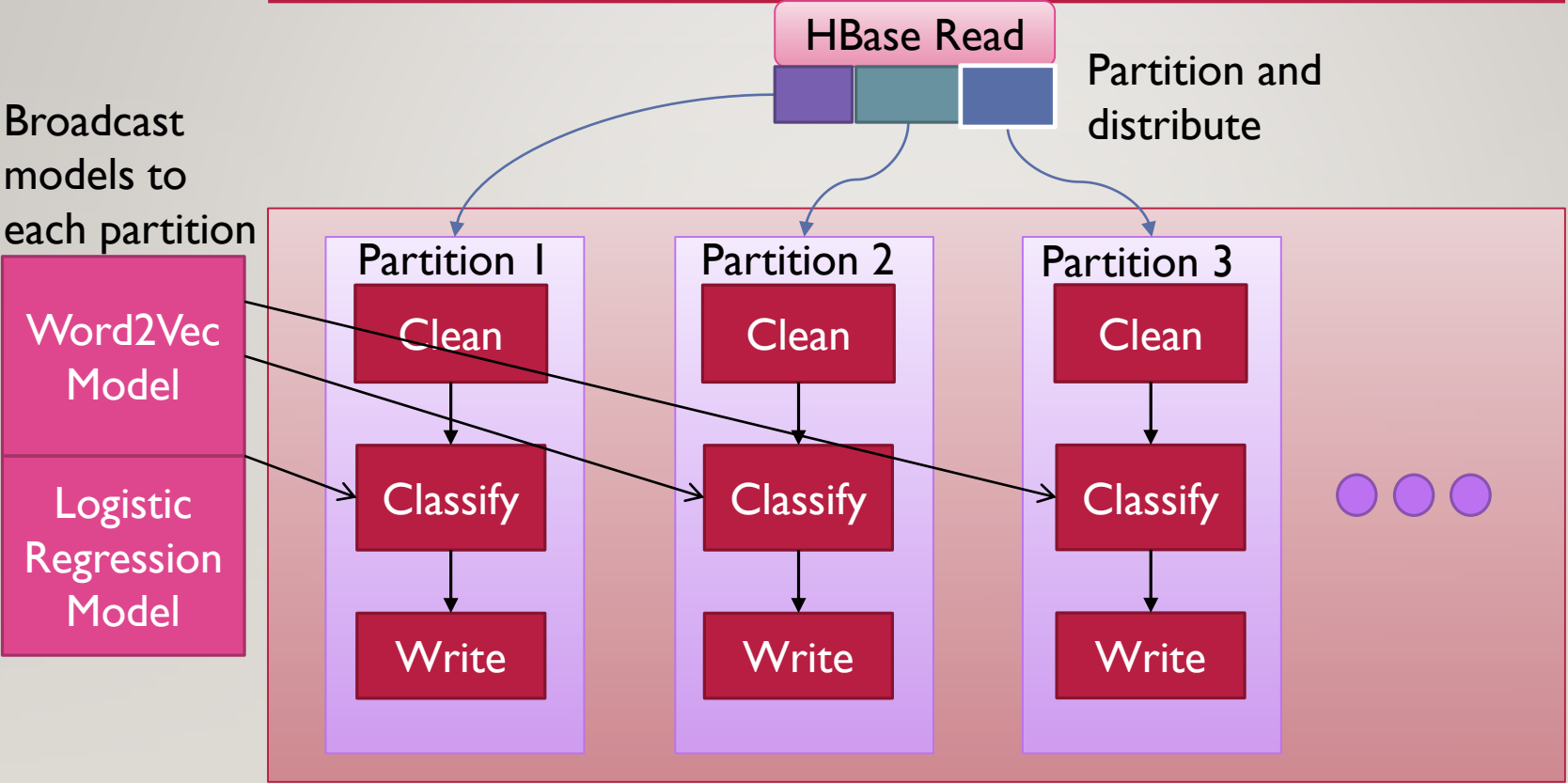
---

- Need to be able to handle large collections efficiently to classify all of the tweets
- Classify at a rate faster than the tweets coming in
- Allow reruns as more classes are added to the training set

# CLASSIFICATION RUNTIME PERFORMANCE

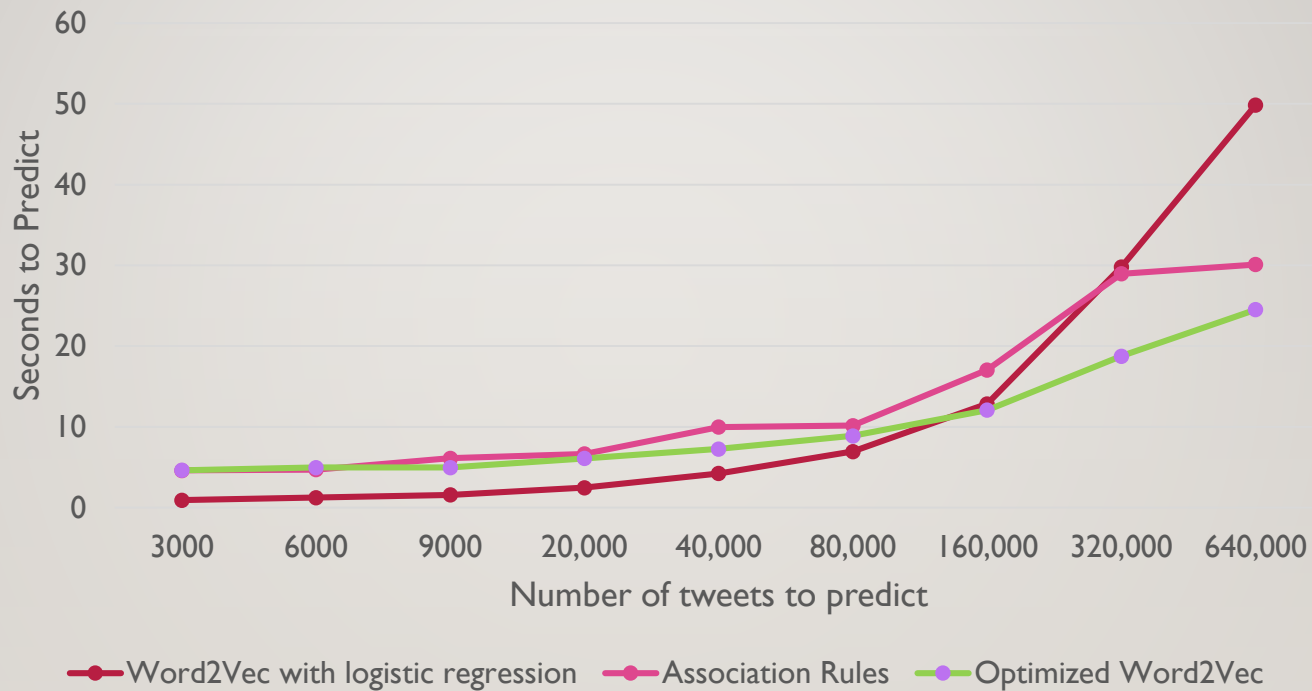


# OPTIMIZATION





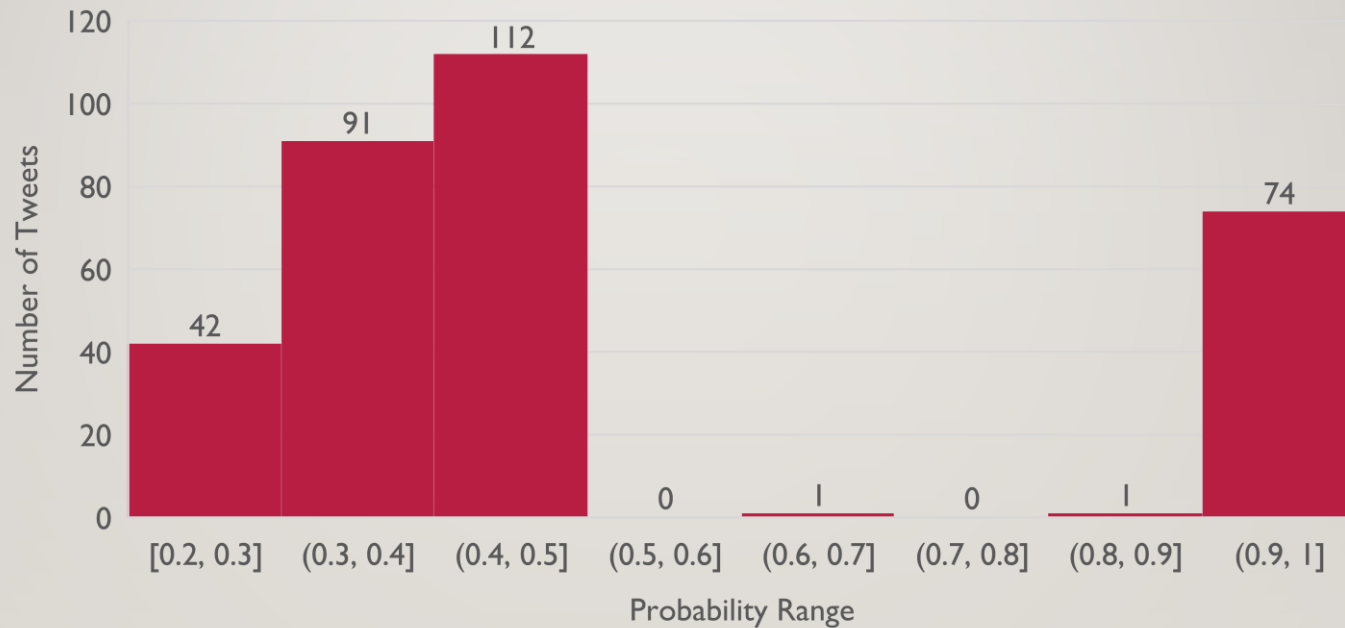
# PROCESSING ACROSS PARTITIONS INCREASES RUNTIME PERFORMANCE!



57% faster than original Word2Vec  
14% faster than Association Rules

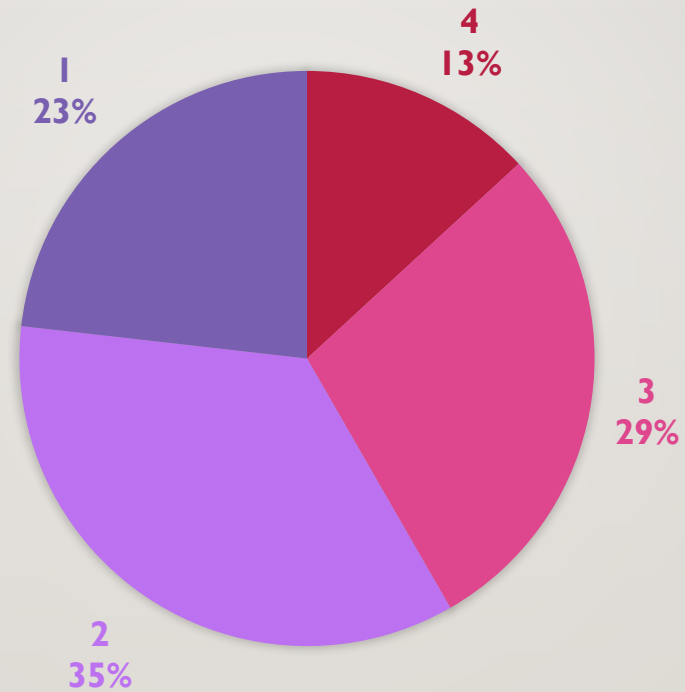
# PROBABILITY DISTRIBUTION FOR TEST DATA

---



# MULTI-CLASS ASSIGNMENT DISTRIBUTION FOR TEST DATA

---



# CONCLUSION

---

- Reading data in blocks from HBase and then partitioning it into parallel tasks results in huge run time performance efficiency and predictability.
- Cleaning text based on the English usage nuances in the Twitter universe results in better accuracy.
- Feature selection methods like Word2Vec that capture richer word semantics and context result in better accuracy than traditional ones for text classification.
- It is natural for a Tweet to be classified in multiple classes and the tradeoff between precision and recall is dependent on the user/product requirements.

# FUTURE WORK

---

- The system can be retrained using a bigger corpus to generate a newer set of word vectors. Training on a text corpus like Google News can help generate word vectors that have richer word relationships encoded within. These can help improve the classification accuracy.
- The Logistic Regression classifier can be retrained on new classes.
- The system will be configured to run via a cron job periodically.
- In addition to classifying a tweet, the system also emits probabilities of all the classes that could be saved in HBase and can be used by SOLR or the front-end team to use as a criterion for customizing the indexing or user experience.
- Comparisons can be performed with the results of the developed classifier with the AR classifier or a few more classifiers and an inter-classifier agreement analysis can throw further light on the efficacy of the developed classifier.

# ACKNOWLEDGEMENTS

---

We would like to acknowledge and thank the following for assisting and supporting us throughout this project.

- Dr. Edward Fox, Dr. Denilson Alves Pereira
- NSF grant IIS - 1619028, III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR)
- NSF grant IIS - 1319578, III: Small: Integrated Digital Event Archiving and Library (IDEAL)
- Digital Library Research Laboratory
- Graduate Research Assistant – Sunshin Lee
- Other teams in CS 5604



**KEEP  
CALM  
PRESENTATION IS OVER  
ANY  
QUESTIONS?**

# APPENDIX

---

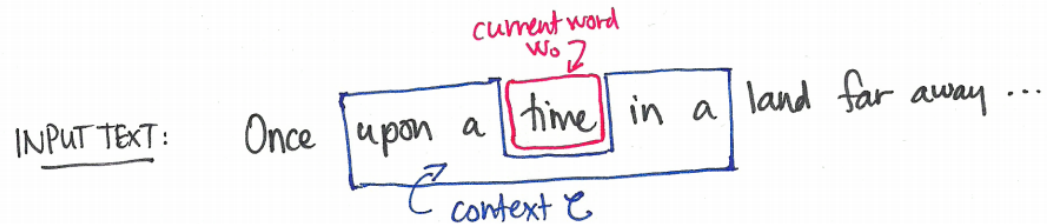




# FEATURE SELECTION TECHNIQUES

Technique	Advantages	Disadvantages
Tf-idf	Superior for small feature sets that have a large scatter of features among the classes. High term removal capability.	Accuracy suffers for large data sets where a term distribution alone does not suffice in class discrimination.
Mutual information	Simple to implement.	Inferior performance in estimation of probabilities because of bias.
Association rules	Fast execution. Very good accuracy for multi-class scenarios. Rule based classifier helps understand the classification decision easily.	Prone to discovering too many rules or poorly understandable rules that hurt performance and interpretation.
Chi-square statistic	Robust accuracy and performance with large sample sets with fewer classes.	Difficulty in interpretation of when there are a large number of classes.
Within class popularity	Identifies words that are most discriminative.	Ignores the sequence of words.
Word2Vec	Captures relationships of a word with neighbors.	High computation complexity. Long training time for large sample size.

# WORD2VEC



sample #	$w_0$	context $C$
1	once	{upon, a}
	...	
4	time	{upon, a, in, a}
	...	

# WORD2VEC

---

- Can learn the word vectors via two forms.

- CBOW

Given only the current context  $\mathcal{C}$ , e.g.

$$\mathcal{C} = \{\text{upon, a, in, a}\}$$

predict which of all possible words is the current word  $w_0$ , e.g.

$w_0 = \text{time}$ .

Predict the word, given the context.

# WORD2VEC

---

- Skip-gram – Inverse objective of CBOW. Predict the context, given a word.

# REAL WORLD EVENTS USED FOR EXPERIMENTS

---

The real world event along with the amount of tweets labeled as that class for our experimental sets.

Hurricane Sandy (108 tweets)	Hurricane Isaac (83 tweets)
New York Firefighter Shooting (58 tweets)	Kentucky Accidental Child Shooting (16 tweets)
Newtown School Shooting (157 tweets)	Manhattan Building Explosion (189 tweets)
China Factory Explosion (178 tweets)	Texas Fertilizer Explosion (120 tweets)
Hurricane Arthur (169 tweets)	

# REAL WORLD EVENTS CLASSIFIED

The real world events classified along with some collections tweets of that event are found.

Real World Event	Collections	Real World Event	Collections
Hurricane Sandy	23,27,375	Hurricane Isaac	27,28,375
New York Firefighter Shooting	43,46	Kentucky Accidental Child Shooting	45,46
Newtown School Shooting	41,42,46	Manhattan Building Explosion	173,174,399,400
China Factory Explosion	231,232	Texas Fertilizer Explosion	77,381
Hurricane Arthur	27,187,188,375	Quebec Train Derailment	96,98,381
Fairdale Tornado	406,632	Oklahoma Tornado	406,84
Mississippi Tornado	406,528	Alabama Tornado	406,407