Sentiment analysis of IMDb movie reviews

By : - Group 28
Paritosh Gaiwak (pgaiwak)
Pragam Gandhi (pmgandh2)
Sanjana Kacholia (skachol)
Tushar Pahuja (tpahuja)

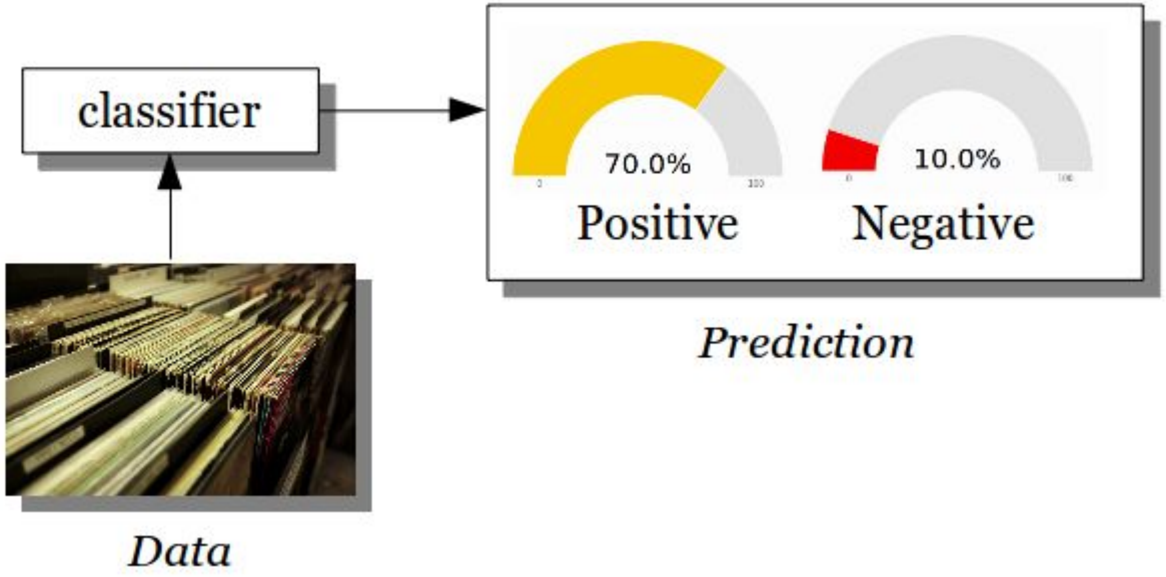
Contents

- Sentiment analysis
- Dataset
- Approach
- Algorithms
- Comparison
- Conclusion
- References

Sentiment analysis

- Understanding sentiment from language data
- Subfield of NLP
- Applications:
  - Recommender systems
  - Movie performance evaluation

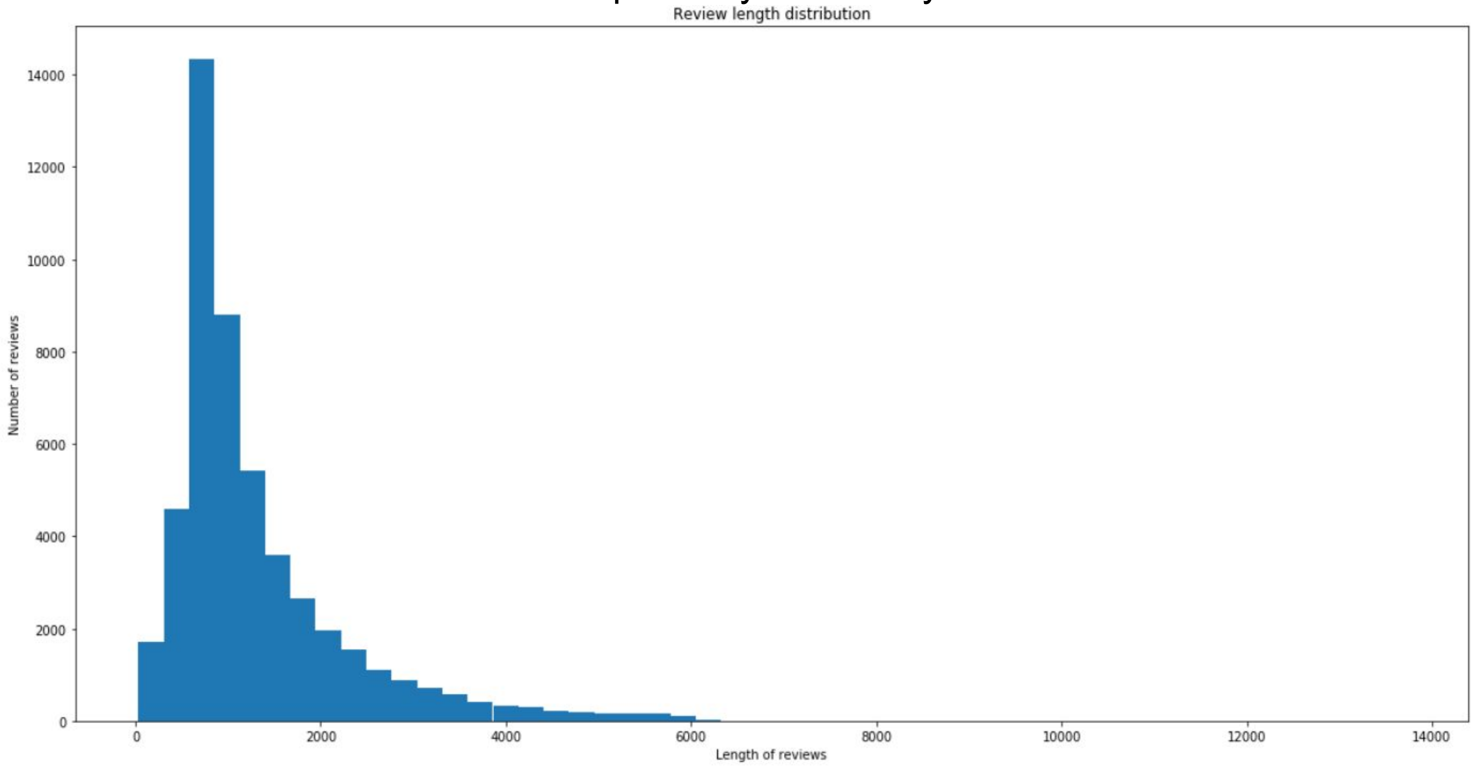
Importance: Derive structure from unstructured data



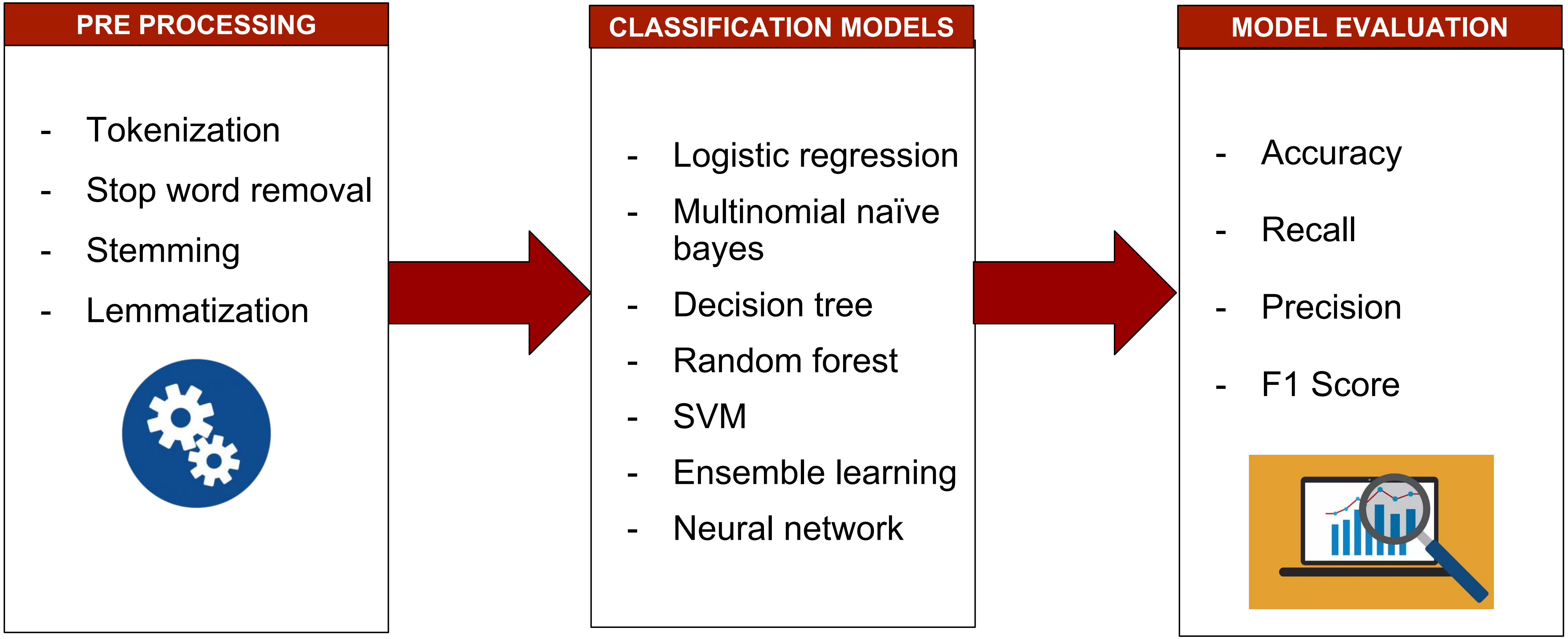
Dataset

- 50000 samples: 25000 positive, 25000 negative (binary)
- Random sampling
- Preprocessing:
  - Tokenization
  - Stop word removal
  - Stemming
  - Lemmatization

Exploratory data analysis



Approach



PRE PROCESSING

- Tokenization
- Stop word removal
- Stemming
- Lemmatization

CLASSIFICATION MODELS

- Logistic regression
- Multinomial naïve bayes
- Decision tree
- Random forest
- SVM
- Ensemble learning
- Neural network

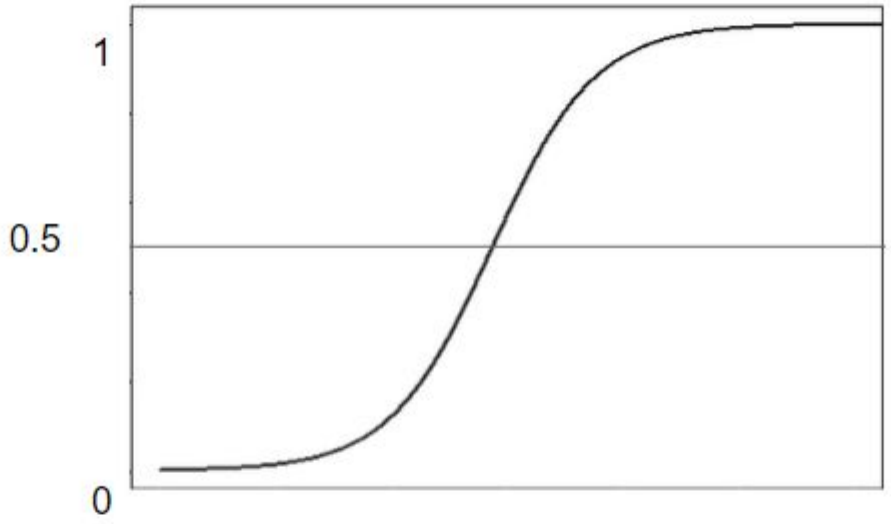
MODEL EVALUATION

- Accuracy
- Recall
- Precision
- F1 Score

Logistic regression

- Models probability of default class
- Go-to method for binary classification

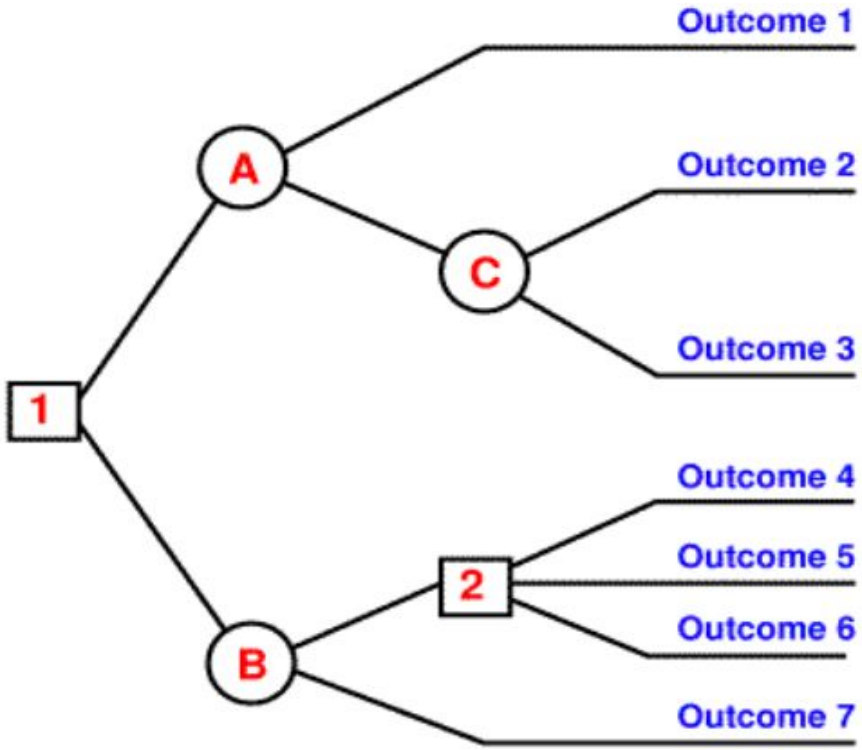
Precision	**0.879**
Recall	**0.9**
Accuracy	**0.89**
F1	**0.89**



Decision tree

- Classification using tree; leaves: class labels; branches: features

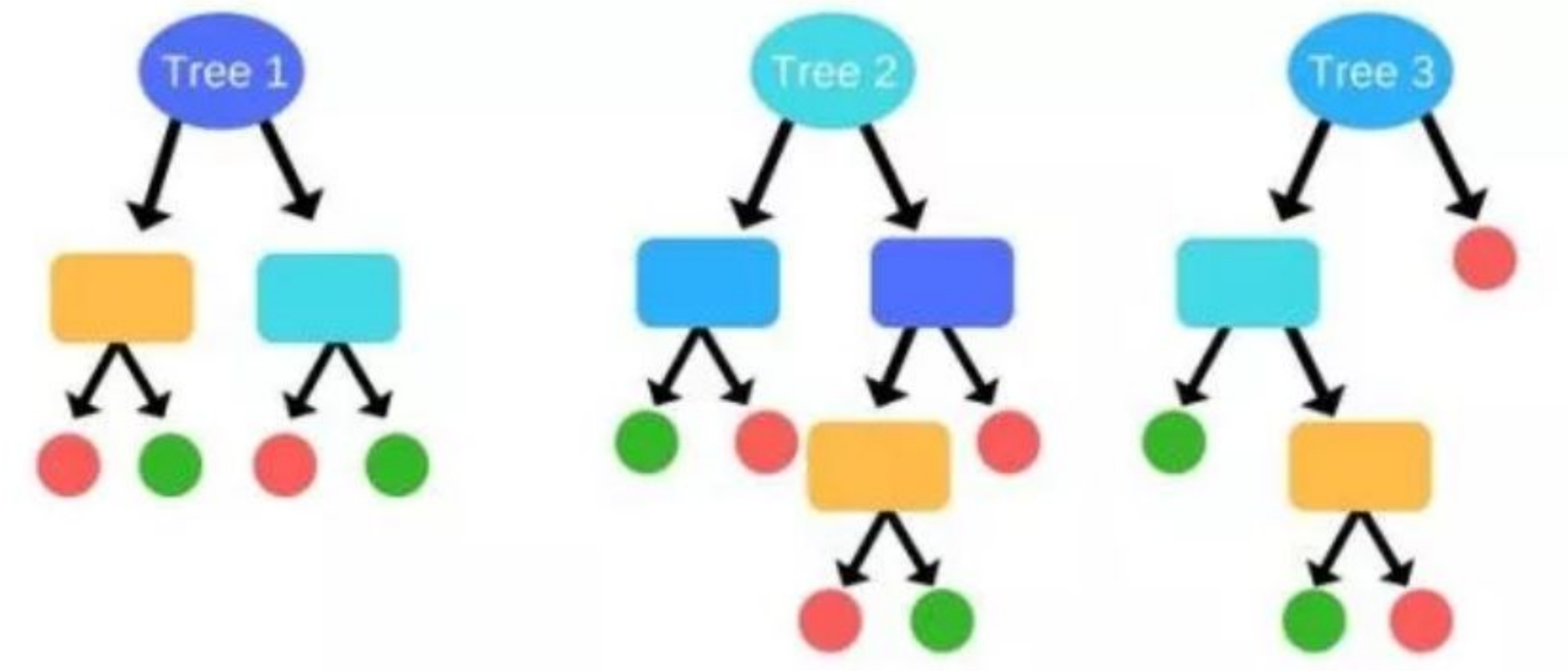
Precision	**0.72**
Recall	**0.71**
Accuracy	**0.71**
F1	**0.72**



Random forest

- Ensemble learning classification (multiple decision trees)

Precision	**0.85**
Recall	**0.84**
Accuracy	**0.84**
F1	**0.85**



Multinomial naïve bayes

- Classification with discrete features (e.g. word counts for text classification)

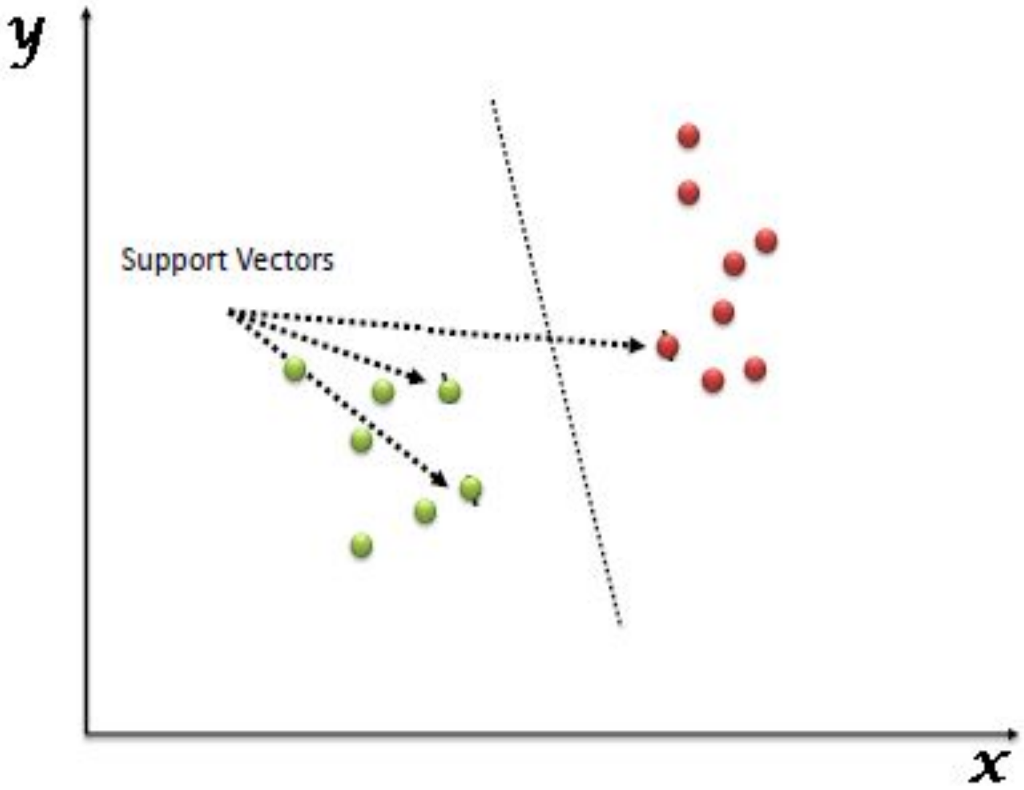
Precision	**0.85**
Recall	**0.87**
Accuracy	**0.86**
F1	**0.86**

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Support vector classifier

- Supervised machine learning algorithm
- Kernel trick transforms data; finds optimal boundary between outputs

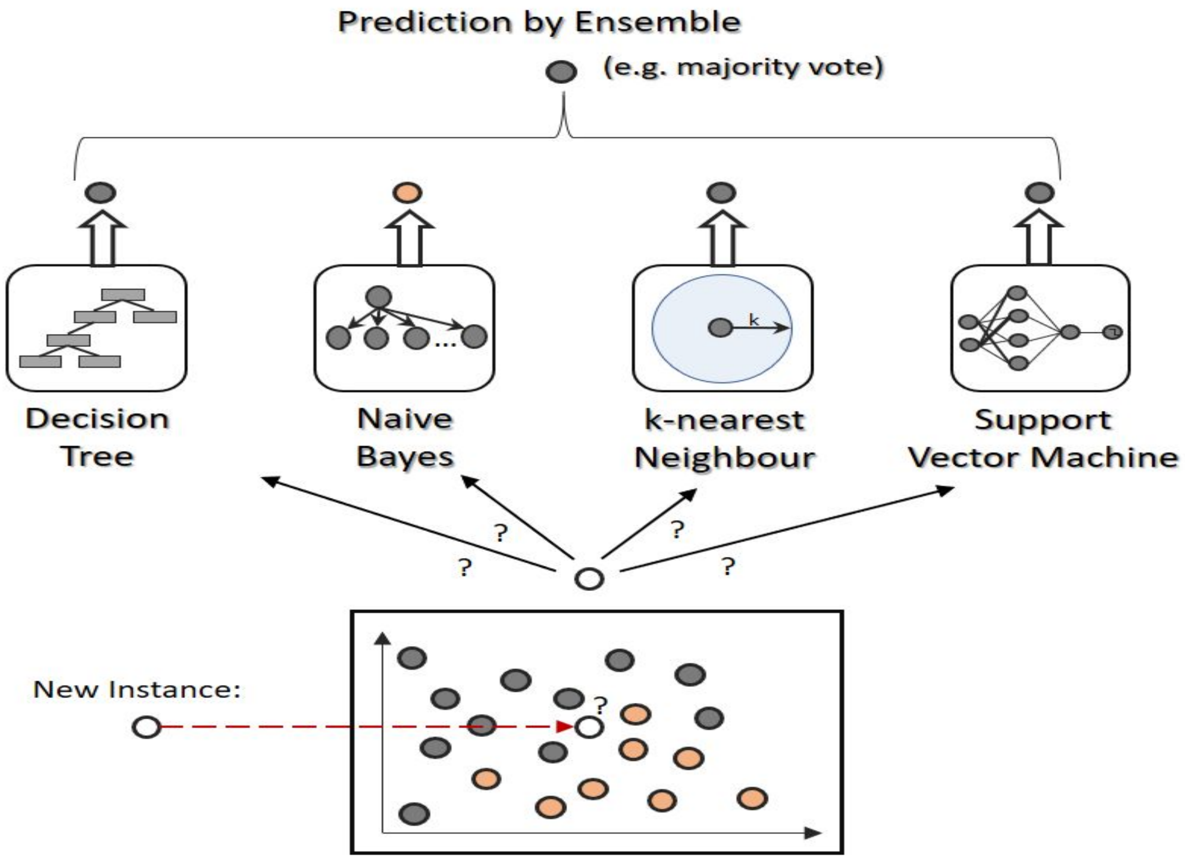
Precision	0.88
Recall	0.89
Accuracy	0.89
F1	0.88



Ensemble learning (different weights)

- Each classifier given different weight

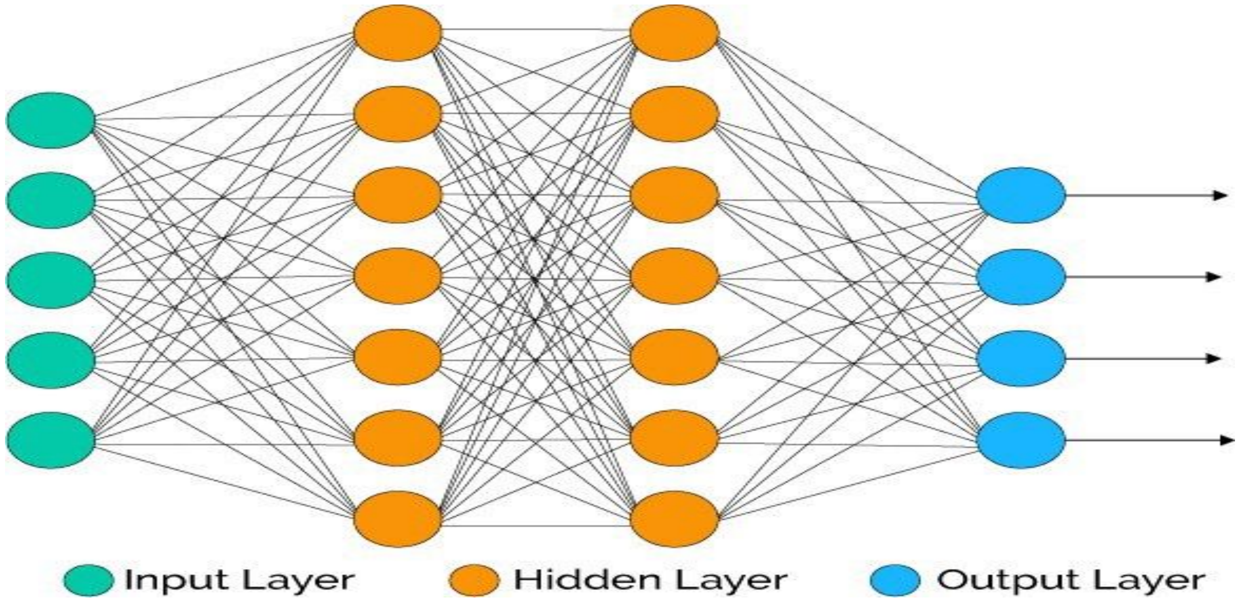
Precision	**0.87**
Recall	**0.9**
Accuracy	**0.89**
F1	**0.89**



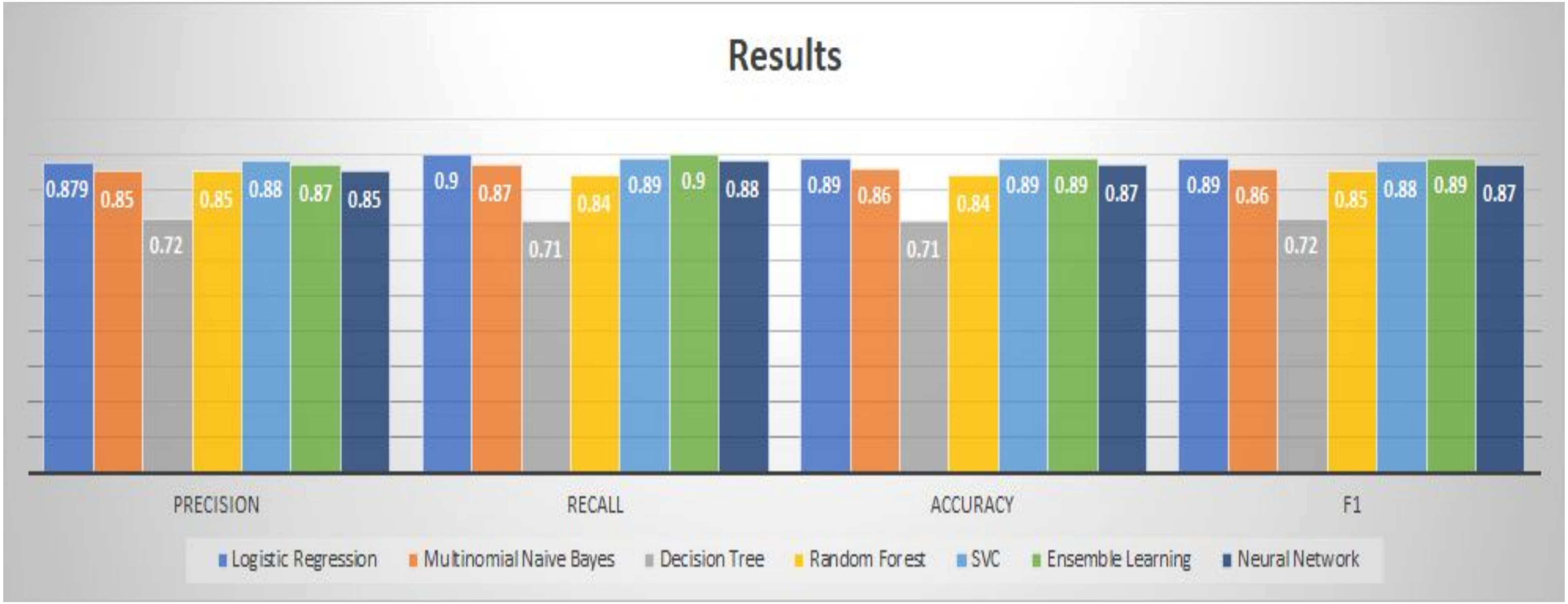
Neural network

- Artificial neural network with multiple hidden layers

Precision	**0.879**
Recall	**0.9**
Accuracy	**0.89**
F1	**0.89**



Comparative analysis



Conclusion

- Models based on accuracy: Logistic > SVC > Ensemble
- Decision tree performs worst

References

- Describes both symbolic and machine learning techniques for understanding sentiments from the text [https://ieeexplore.ieee.org/document/6726818]

- Baselines and bigrams: Simple, good sentiment and topic classification [https://dl.acm.org/citation.cfm?id=2390688]