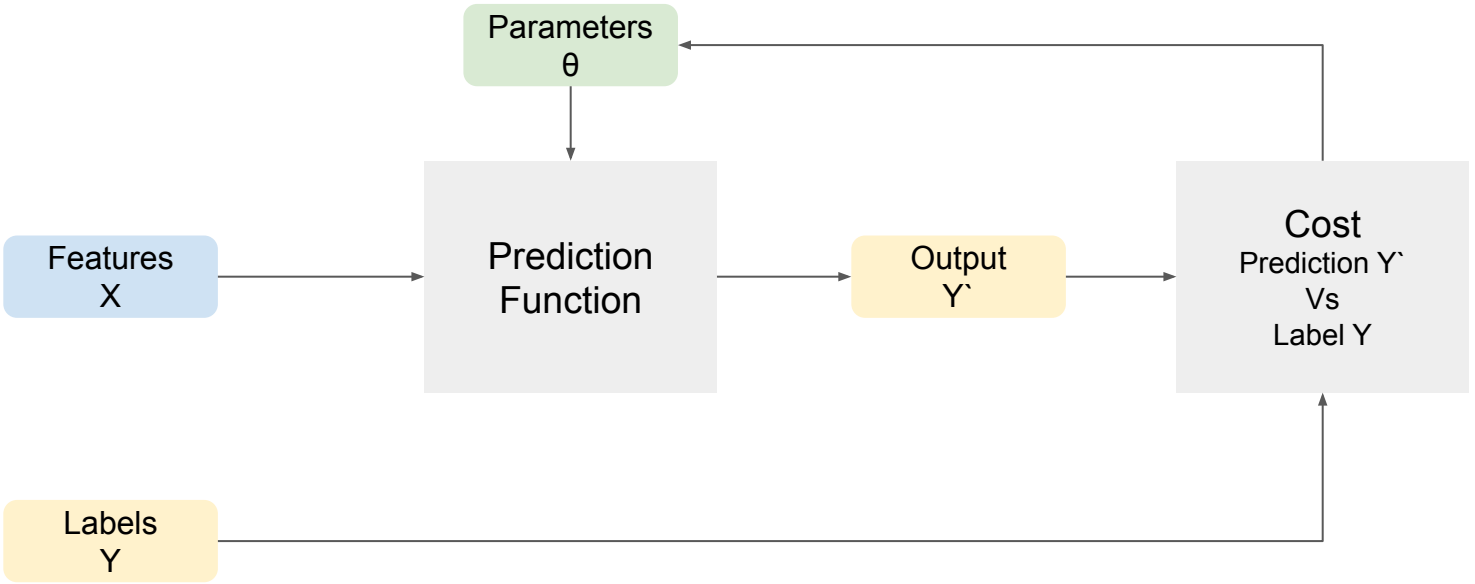
# NLP

## Week - 1

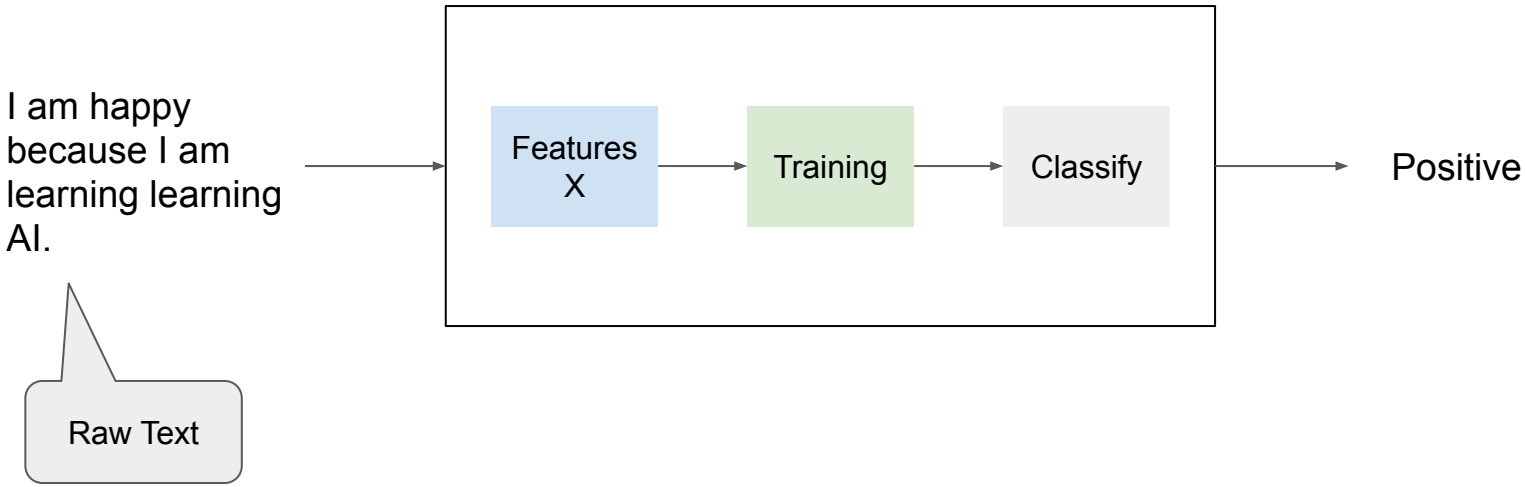
Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Supervised Learning



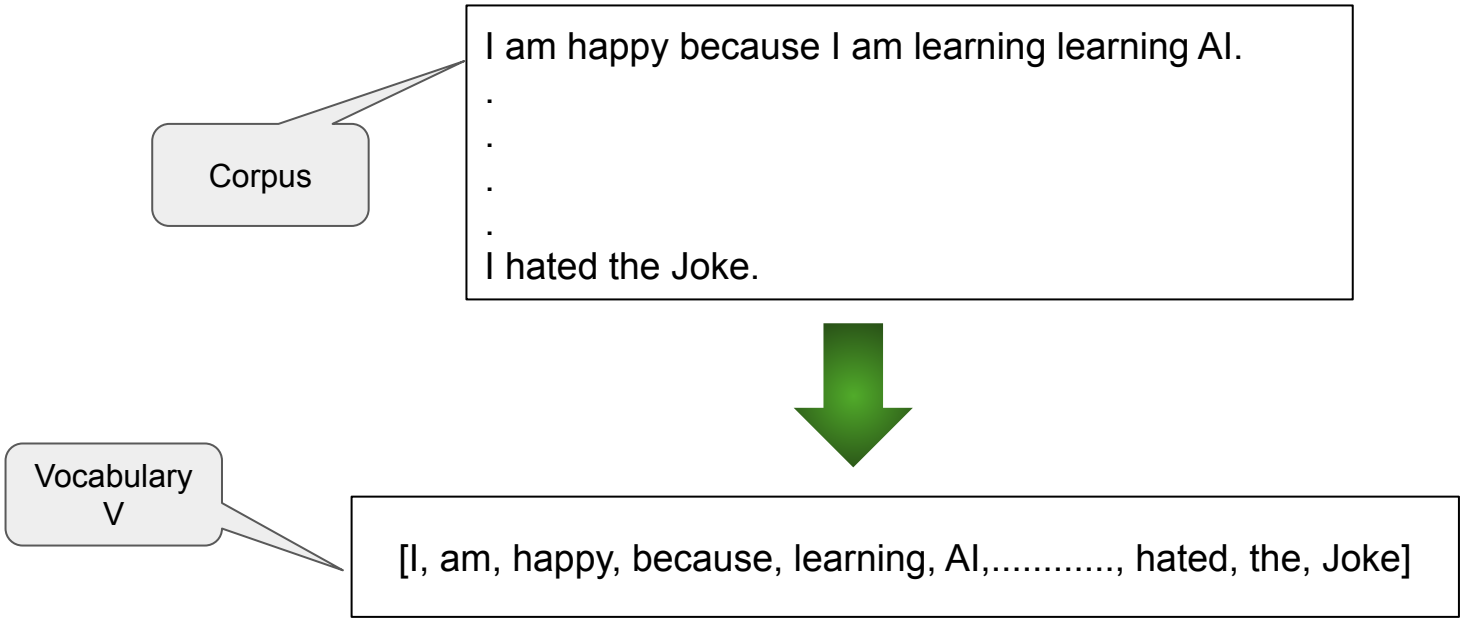

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Sentiment Analysis



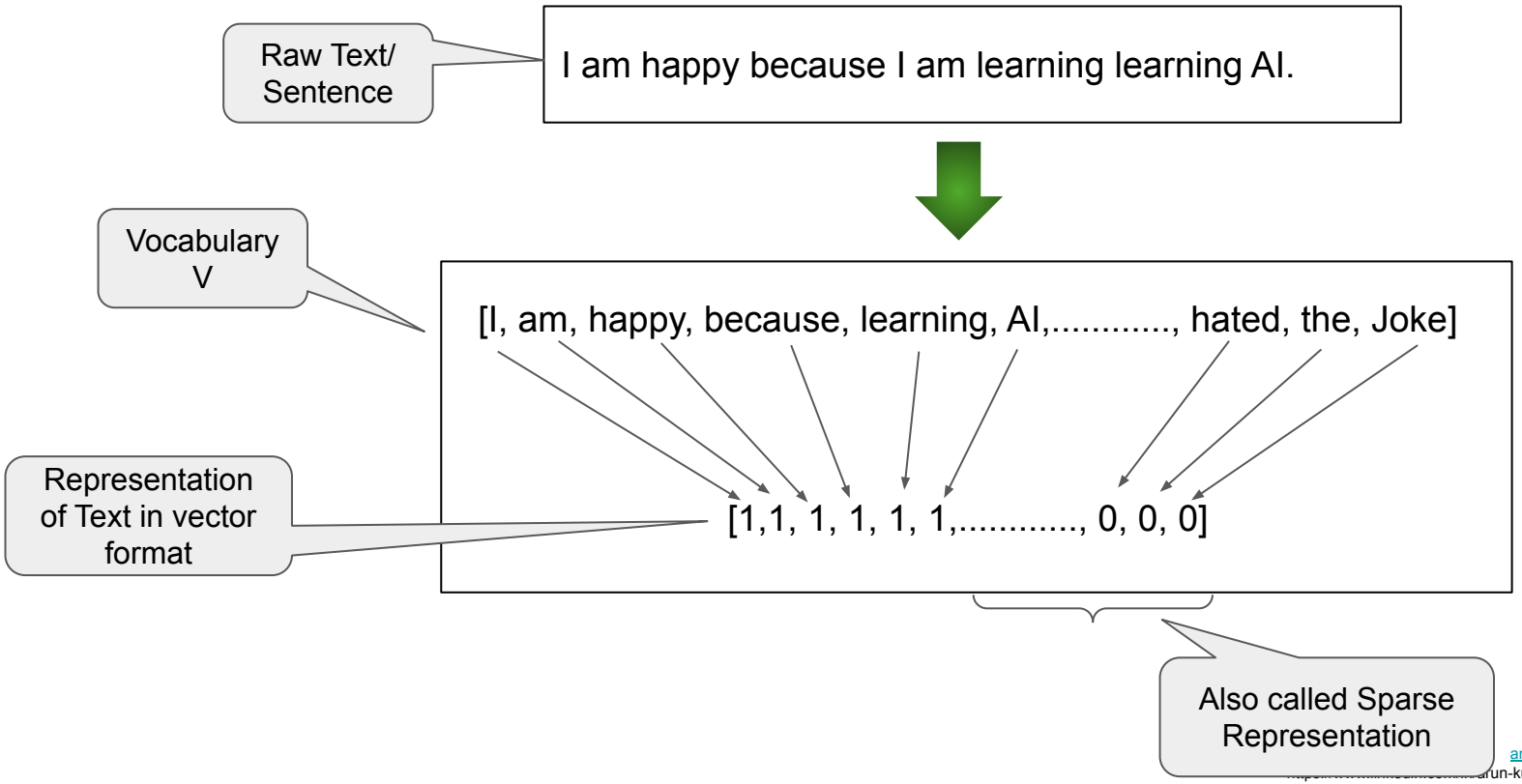

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Vocabulary




Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Feature Extraction




Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Sparse Representation Limitation

I am happy because I am learning learning AI.

[I, am, happy, because, learning, AI,....., hated, the, Joke]

[1,1, 1, 1, 1, 1,....., 0, 0, 0]

$[\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5,.....,\theta_{n-2}, \theta_{n-1}, \theta_n]$

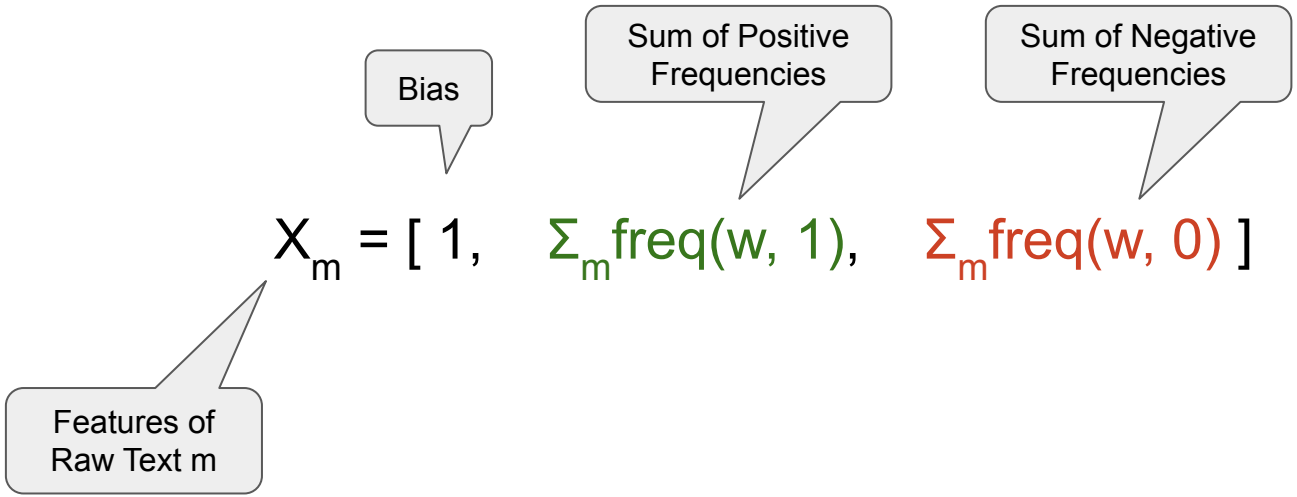
Bias

Too Many Zeros.
Has to find out (Vn + 1) parameters.

- Large Training time
- Large Prediction time

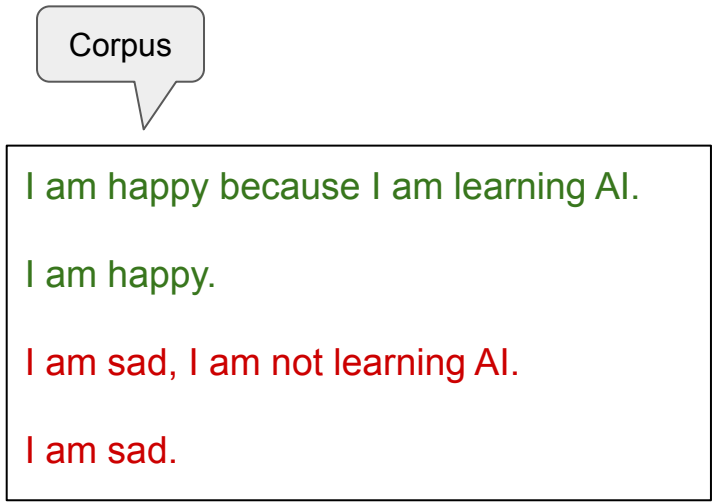
Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

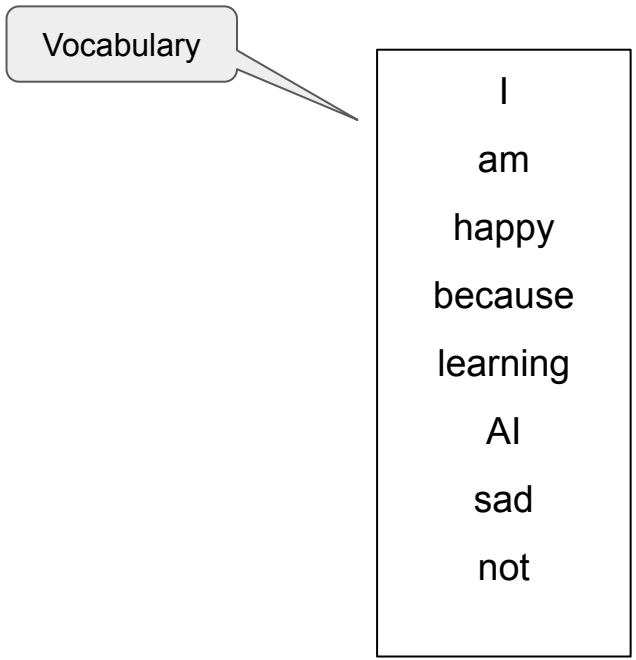
# Another Approach for Feature Extraction




Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Sentiment Analysis




Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Sentiment Analysis - Word Frequency in Classes

Positive Corpus

I am happy because I am learning AI.

I am happy.

| Vocabulary | Frequency |
|---|---|
| I | 3 |
| am | 3 |
| happy | 2 |
| because | 1 |
| learning | 1 |
| AI | 1 |
| sad | 0 |
| not | 0 |

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Sentiment Analysis - Word Frequency in Classes

Negative Corpus

I am sad, I am not learning AI.

I am sad.

| Vocabulary | Frequency |
|---|---|
| I | 3 |
| am | 3 |
| happy | 0 |
| because | 0 |
| learning | 1 |
| AI | 1 |
| sad | 2 |
| not | 1 |

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Sentiment Analysis - Word Frequency in Classes

| Vocabulary | Positive Frequency | Negative Frequency |
|---|---|---|
| I | 3 | 3 |
| am | 3 | 3 |
| happy | 2 | 0 |
| because | 1 | 0 |
| learning | 1 | 1 |
| AI | 1 | 1 |
| sad | 0 | 2 |
| not | 0 | 1 |

This is called
**Frequency Dictionary Mapping**
from (word, class) : Frequency

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Sentiment Analysis - Feature Extraction

| Vocabulary | Positive Frequency | Negative Frequency |
|---|---|---|
| I | 3 | 3 |
| am | 3 | 3 |
| happy | 2 | 0 |
| because | 1 | 0 |
| learning | 1 | 1 |
| AI | 1 | 1 |
| sad | 0 | 2 |
| not | 0 | 1 |

Raw Text

I am sad, I am not learning AI.

$X_m = [\ 1, \quad \Sigma_m freq(w, 1), \quad \Sigma_m freq(w, 0)\ ]$

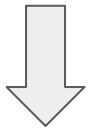
8 11

$X_m = [\ 1, \quad 8, \quad 11]$

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Preprocessing : Stop words, Punctuation, Handles, URLs

@Arun @Kabir
Travelling Made Fun. Travel While Making Friends.
#ridemate #travelmate #bengaluru #delhi #newdelhi



Travelling Made Fun. Travel While Making Friends.

| Stop Words |
| --- |
| and |
| is |
| are |
| at |
| has |
| for |
| a |
| it |

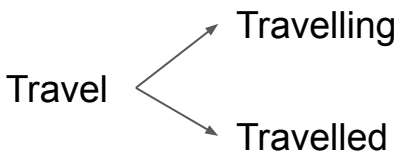
| Punctuation |
| --- |
| , |
| . |
| ; |
| : |
| ‘ |
| ] |
| [ |
| ! |

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

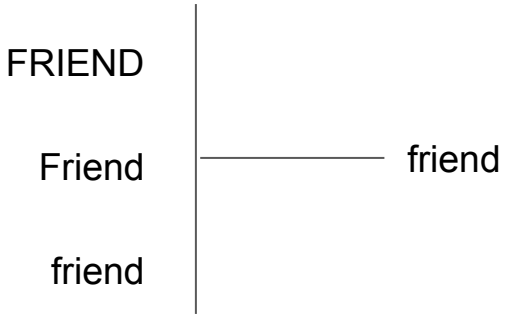
# Preprocessing : Stemming and Lowercase

**Stemming**

Travelling Made Fun. Travel While Making Friends.



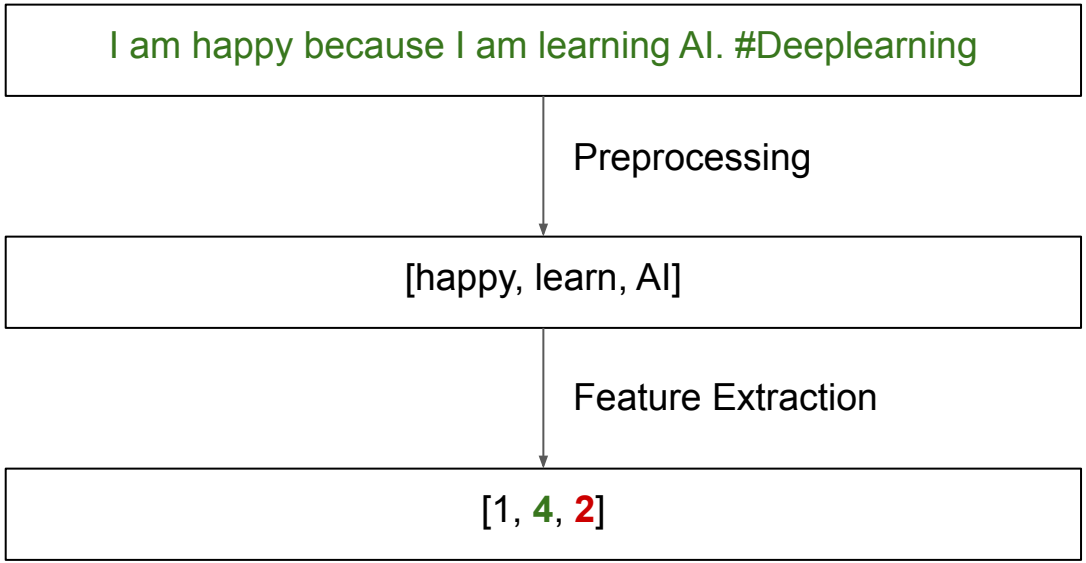

**Lowercase**




| Processed Text | [travel, made, fun, while, making, friend] |
|---|---|

un Kumar Anala
un.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

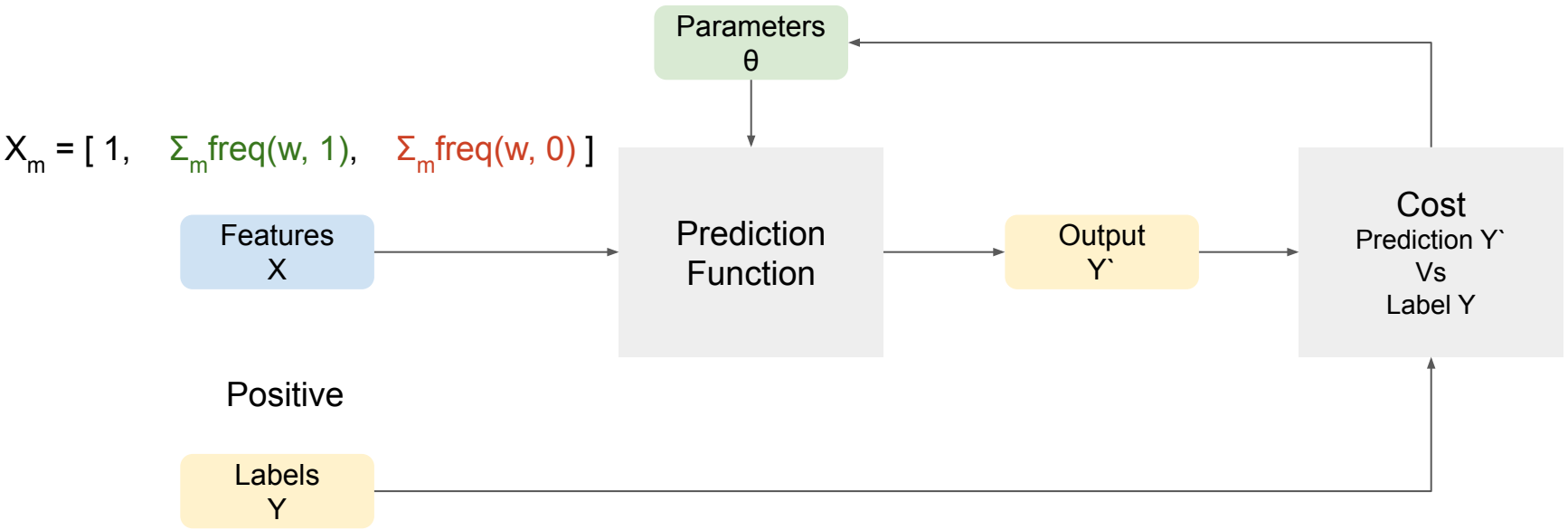
# General overview




$X_m = [\ 1, \quad \Sigma_m freq(w, 1), \quad \Sigma_m freq(w, 0)\ ]$

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

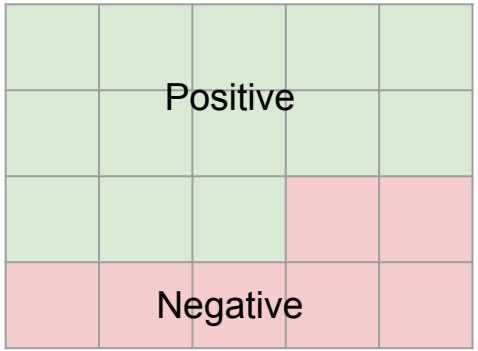
# Supervised Learning



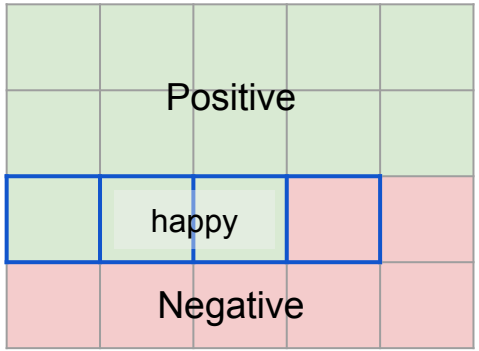

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Probability and Bayes Rule

Corpus of texts



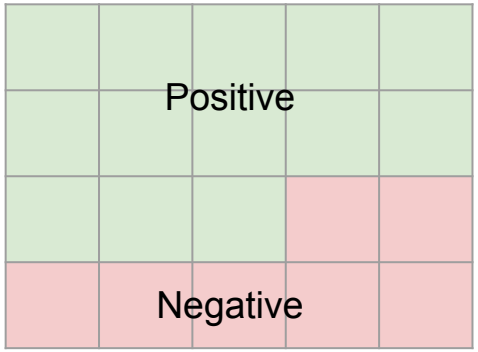

Texts containing the word happy




Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Probability and Bayes Rule

Corpus of texts



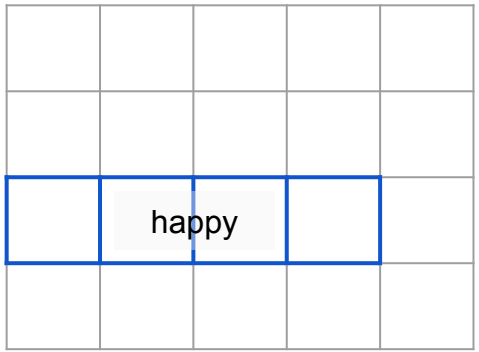

P(positive) = $N_{pos}$ / N = 13/20 = 0.65

P(negative) = 1 - 0.65 = 0.35

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Probability and Bayes Rule

Corpus of texts

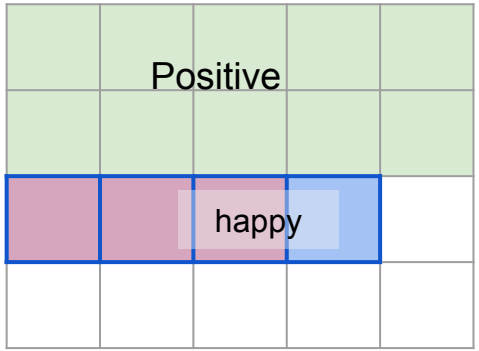


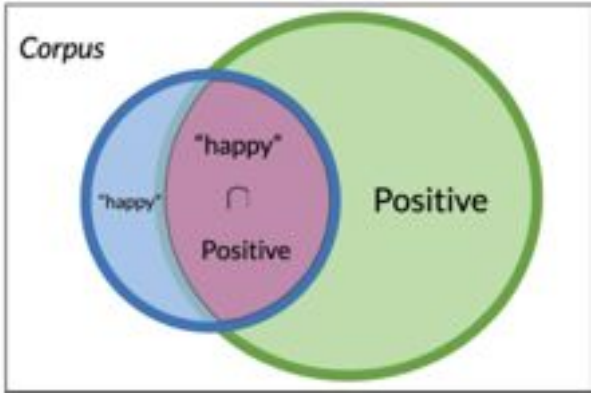

B → Texts contains “happy”

P(B) = P(happy) = 4 / 20 = 0.2

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Probability of Intersection

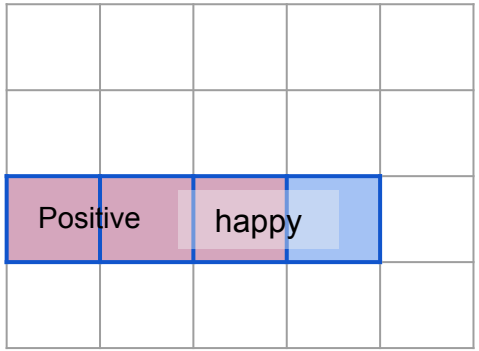


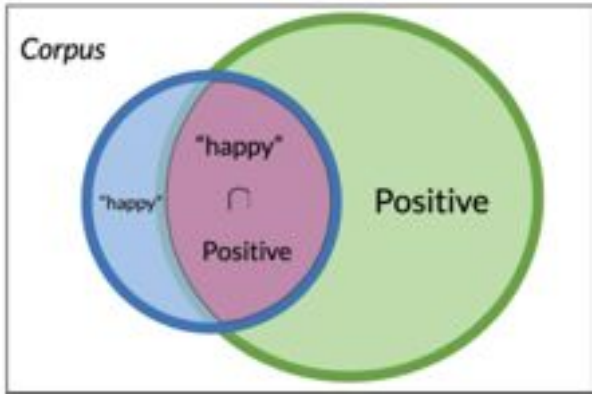




$P(A \cap B) = P(A, B) = 3 / 20 = 0.15$

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Conditional Probability




P(A|B) = P(Positive|happy)

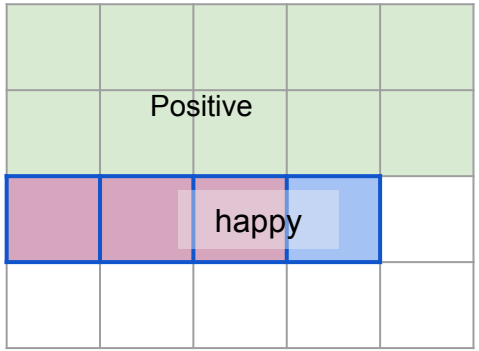
= 3 / 4 = 0.75

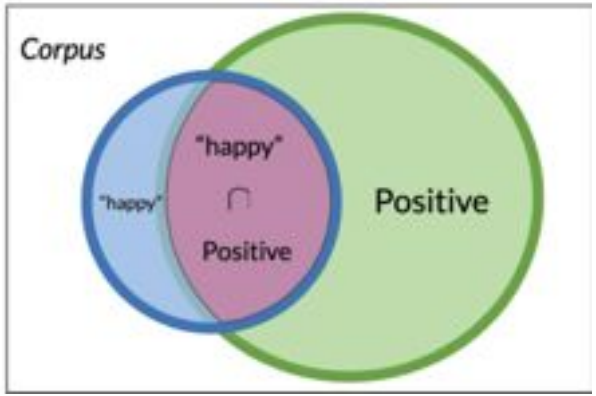
Text has 75% likelihood of being positive,

if it contains the word “happy”

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Conditional Probability




P(B|A) = P(happy|positive)

= 3 / 13 = 0.231

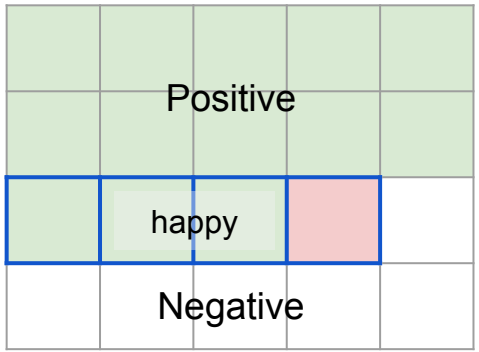
Text has 23.1% likelihood of containing the word “happy”, if the text is positive.

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Bayes Rule

P(Positive | “happy”) = P(Positive ∩ “happy”) / P(“happy”)

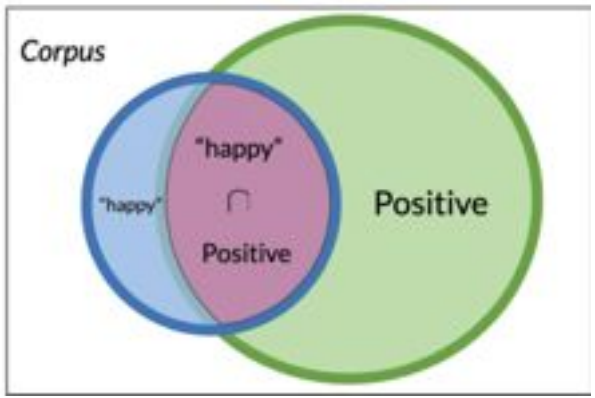
P(“happy” | Positive) = P(Positive ∩ “happy”) / P(Positive)




Bayes’ rule

$$P(\text{Positive}|\text{“happy”}) = P(\text{“happy”}|\text{Positive}) \times \frac{P(\text{Positive})}{P(\text{“happy”})}$$

$$P(X|Y) = P(Y|X) \times \frac{P(X)}{P(Y)}$$




Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Naive Bayes Rule for Sentiment Analysis

Corpus

I am happy because I am learning AI.

I am happy, not sad.

I am sad, I am not learning AI.

I am sad, not happy.

| Vocabulary | Positive Frequency | Negative Frequency |
|---|---|---|
| I | 3 | 3 |
| am | 3 | 3 |
| happy | 2 | 1 |
| because | 1 | 0 |
| learning | 1 | 1 |
| AI | 1 | 1 |
| sad | 1 | 2 |
| not | 1 | 2 |
| | 13 | 12 |

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Naive Bayes Rule for Sentiment Analysis

| Vocabulary | Positive Frequency | Negative Frequency | Positive Probability | Negative Probability |
|---|---|---|---|---|
| I | 3 | 3 | 0.24 | 0.25 |
| am | 3 | 3 | 0.24 | 0.25 |
| happy | 2 | 1 | 0.25 | 0.08 |
| because | 1 | 0 | 0.08 | 0 |
| learning | 1 | 1 | 0.08 | 0.08 |
| AI | 1 | 1 | 0.08 | 0.08 |
| sad | 1 | 2 | 0.08 | 0.17 |
| not | 1 | 2 | 0.08 | 0.17 |
| | 13 | 12 | 1 | 1 |

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Naive Bayes Rule for Sentiment Analysis

I am happy today; I am learning

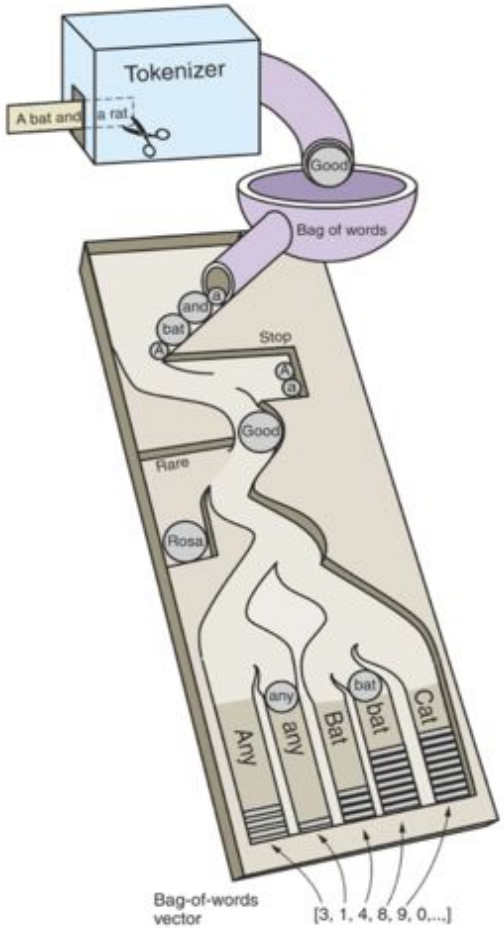
$$\prod_{i=1}^{m} \frac{P(w_i|pos)}{P(w_i|neg)} = \frac{0.14}{0.10} = 1.4$$

$$\frac{0.20}{0.20} * \frac{0.20}{0.20} * \frac{0.14}{0.10} * \frac{0.20}{0.20} * \frac{0.20}{0.20} * \frac{0.10}{0.10}$$

| Vocabulary | Positive Probability | Negative Probability |
|---|---|---|
| I | 0.24 | 0.25 |
| am | 0.24 | 0.25 |
| happy | 0.25 | 0.08 |
| because | 0.08 | 0 |
| learning | 0.08 | 0.08 |
| AI | 0.08 | 0.08 |
| sad | 0.08 | 0.17 |
| not | 0.08 | 0.17 |
| | 1 | 1 |

Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/

# Bag of Words




Arun Kumar Anala
analaarun.k@gmail.com
https://www.linkedin.com/in/arun-kumar-anala-35760523/