NLP PROJECT
MS2

MOHAMED ASHRAF

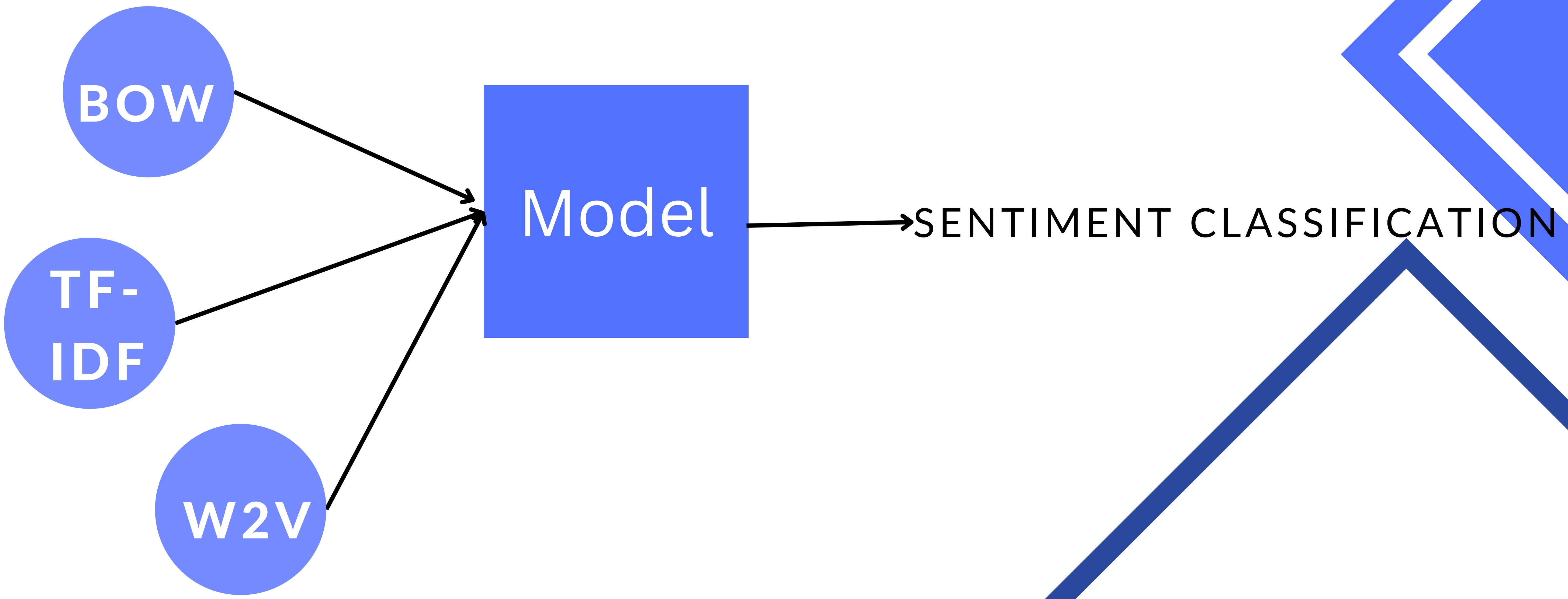
1. DATA SET LIMITATIONS

SMALL SAMPLE SIZE

SAMPLE CHOSEN WAS SMALL SAMPLE COMPARED TO THE PROBLEM WE ARE IMPLEMENTING IT MAY NOT CAPTURE THE FULL DIVERSITY AND COMPLEXITY OF THE PROBLEM DOMAIN. WHICH CAN LEAD TO OVERFITTING OR BIASED RESULTS.

1. METHODOLOGY AND APPROACHES

WE WILL BE TESTING MODELS WITH DIFFERENT FEATURE SETS AND DISCUSS THEIR F SCORES



1. FEATURES PREPARATION

BOW

IN THE FEATURE VECTORS EACH ELEMENT REPRESENTS WHETHER THE CORRESPONDING WORD FROM THE VOCABULARY IS PRESENT (1) OR ABSENT (0) IN THE DOCUMENT.

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

1. FEATURES PREPARATION

TERM FREQUENCY- INVERSE DOCUMENT FREQUENCY

- TERM FREQUENCY REPRESENTS THE FREQUENCY, WHICH MEASURES HOW FREQUENTLY A TERM APPEARS IN A DOCUMENT.
- IDF REPRESENTS THE INVERSE DOCUMENT FREQUENCY, WHICH MEASURES HOW IMPORTANT A TERM IS ACROSS THE ENTIRE COLLECTION OF DOCUMENTS.

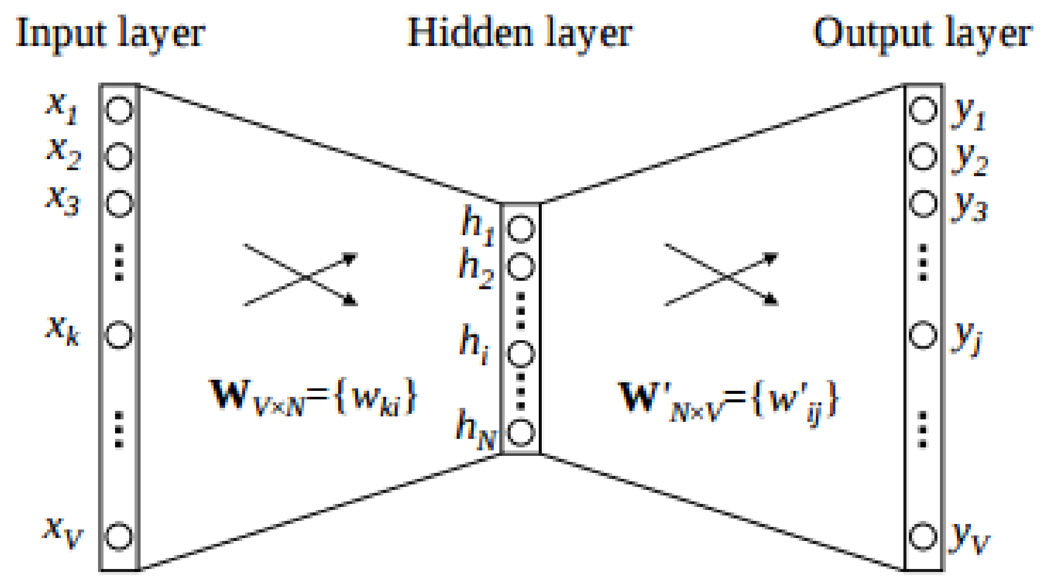
1. WORD2VEC

DENSE VECTORS REPRESENTATION

W2V CAN BE USED TO REPRESENT WORDS AS DENSE VECTORS IN ORDER TO CAPTURE SIMILARITIES BETWEEN THE WORDS USING COSINE SIMILARITY CONCEPT BETWEEN VECTORS

1. WORD2VEC

WORD2VEC MODEL



EACH WORD IS REPRESENTED AS A VECTOR WITH N DIMENSIONS (200) WHICH IS CALCULATED USING W2V SKIPGRAM MODEL AND THE VALUES OF THESE VECTORS IS CALCULATED BY LEARNING ABOUT THE CO OCCURENCE OF THESE WORDS IN DIFFERENT CONTEXTS

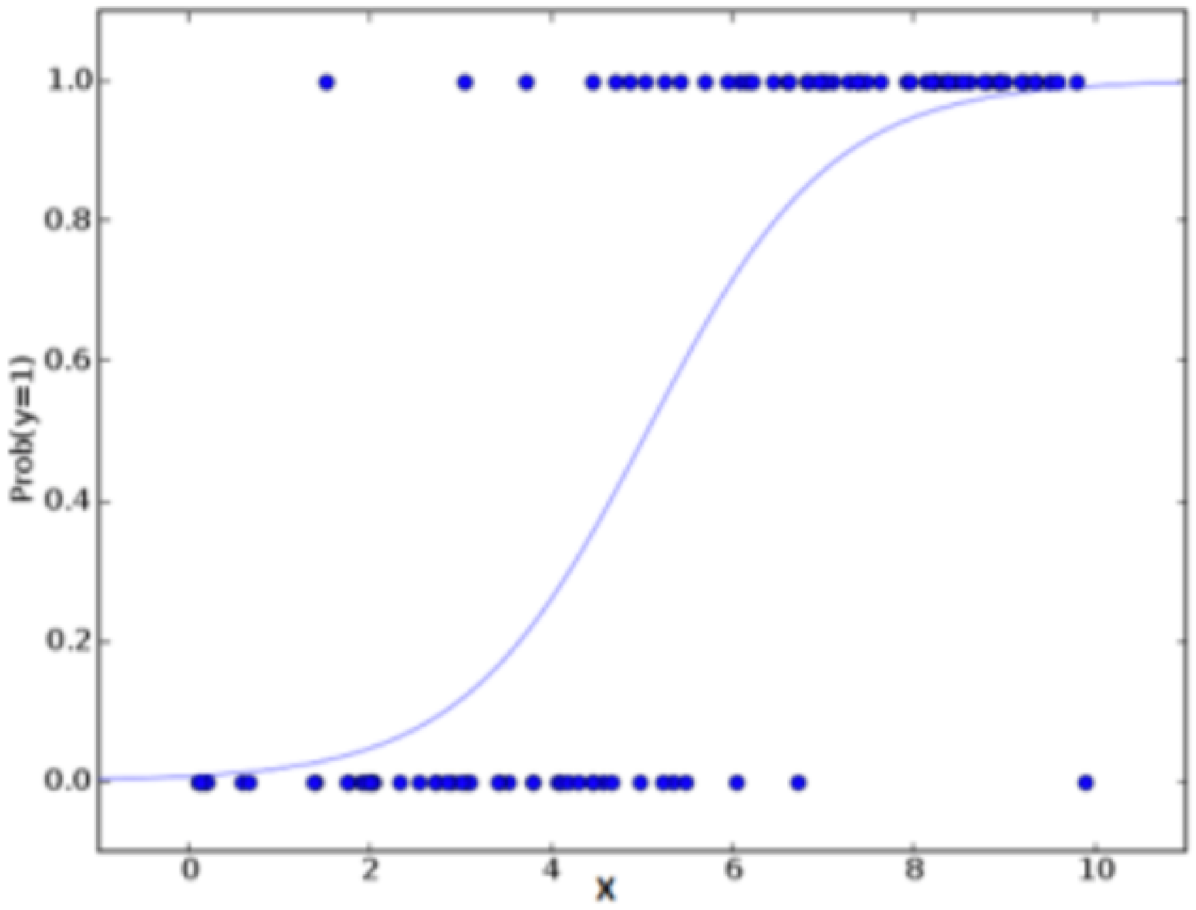
1. MODELS OVERVIEW

IN ORDER TO SOLVE OUR PROBLEM WE HAVE TRIED DIFFERENT MODELS WITH THE FEATURES EXTRACTED IN ORDER TO DETERMINE WHICH MODEL ARCHITECTURE WILL WORK THE BEST SOLVING OUR PROBLEM

1. LOGISTIC REGRESSION MODEL
2. SUPPORT VECTOR MACHINE (SVM)
3. XGBOOST

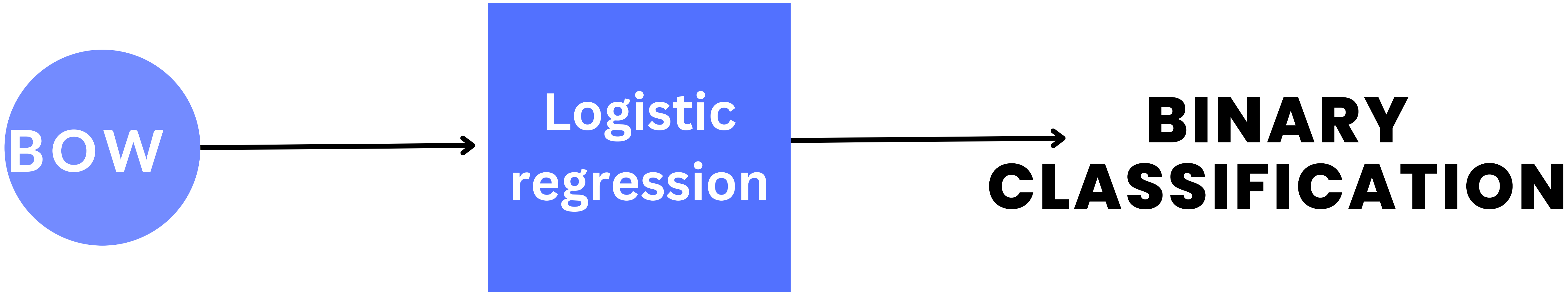
1. LOGISTIC REGRESSION

IN LOGISTIC REGRESSION, THE MODEL LEARNS THE OPTIMAL COEFFICIENTS (WEIGHTS) FOR EACH FEATURE (WORD). IT USES A LOGISTIC FUNCTION (SIGMOID) TO MAP THE INPUT FEATURES TO THE RANGE [0, 1] REPRESENTING THE PROBABILITY OF THE INSTANCE BELONGING TO THE POSITIVE CLASS.



1. LOGISTIC REGRESSION-BOW

1. EXTRACTING TRAIN AND TEST BOW FEATURES
2. TRAINING THE LOGISTIC REGRESSION MODEL
3. PREDICTING ON THE VALIDATION SET
4. CALCULATING THE F1 SCORE FOR THE VALIDATION SET

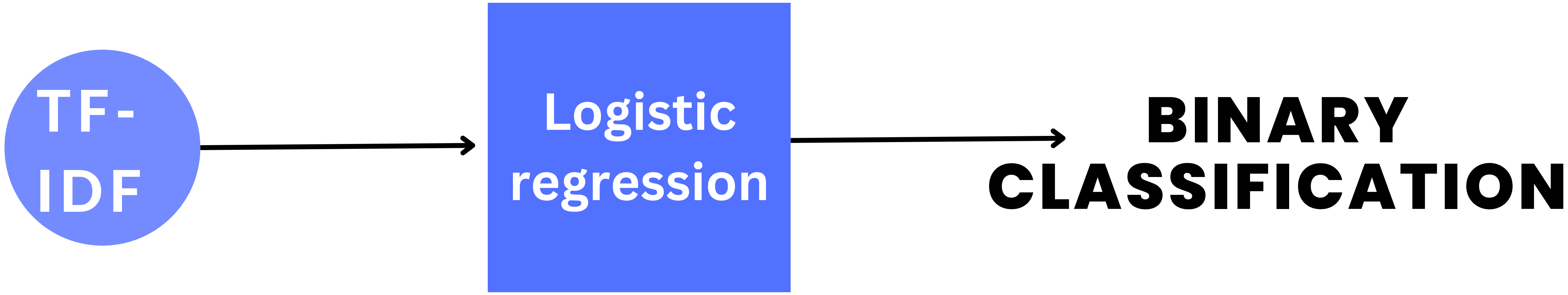



```
mohamed@Mohameds-MacBook-Pro NLP_MS1 % python3 code_1.py
/Users/mohamed/Documents/NLP_MS1/code_1.py:38: FutureWarning: The frame.append method is deprecated
ncat instead.
  tweets_df = train.append(test, ignore_index=True, sort=True)
/Users/mohamed/Documents/NLP_MS1/code_1.py:56: FutureWarning: The default value of regex will change
  tweets_df['clean_tweet'] = tweets_df.clean_tweet.str.replace("[^a-zA-Z#]", " ")
F1 Score for Logistic Regression with BOW Features
0.5303408146300915
```

USING LOGISTIC REGRESSION AND BOW WE OBTAINED ACCURACY OF 0.53

1. LOGISTIC REGRESSION-TFIDF

1. SPLITTING TF-IDF FEATURES
2. TRAINING THE LOGISTIC REGRESSION MODEL
3. PREDICTING ON THE VALIDATION SET
4. CALCULATING THE F1 SCORE FOR THE VALIDATION SET

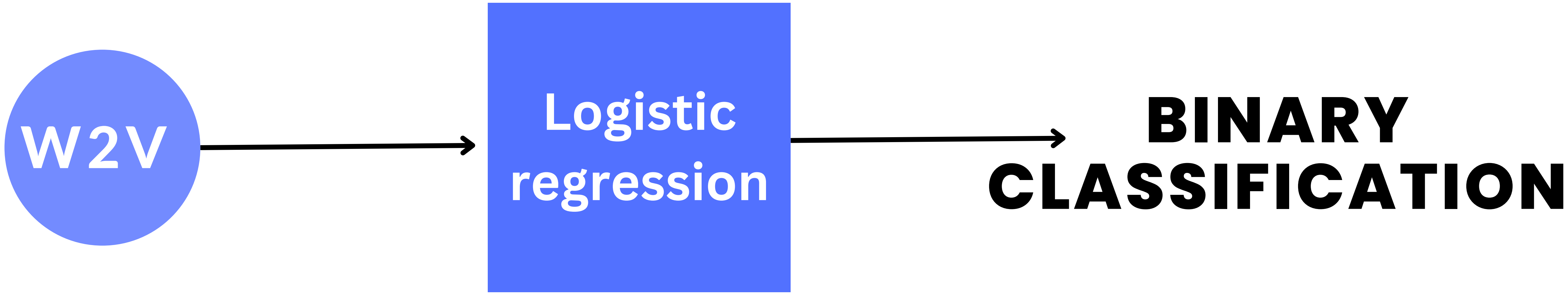



```
mohamed@Mohameds-MacBook-Pro NLP_MS1 % python3 code_1.py
/Users/mohamed/Documents/NLP_MS1/code_1.py:38: FutureWarning: The frame.appe
ncat instead.
  tweets_df = train.append(test, ignore_index=True, sort=True)
/Users/mohamed/Documents/NLP_MS1/code_1.py:56: FutureWarning: The default va
  tweets_df['clean_tweet'] = tweets_df.clean_tweet.str.replace("[^a-zA-Z#]",
0.5451327433628319
```

USING LOGISTIC REGRESSION AND TF-IDF WE OBTAINED ACCURACY OF 0.545

1. LOGISTIC REGRESSION-W2V

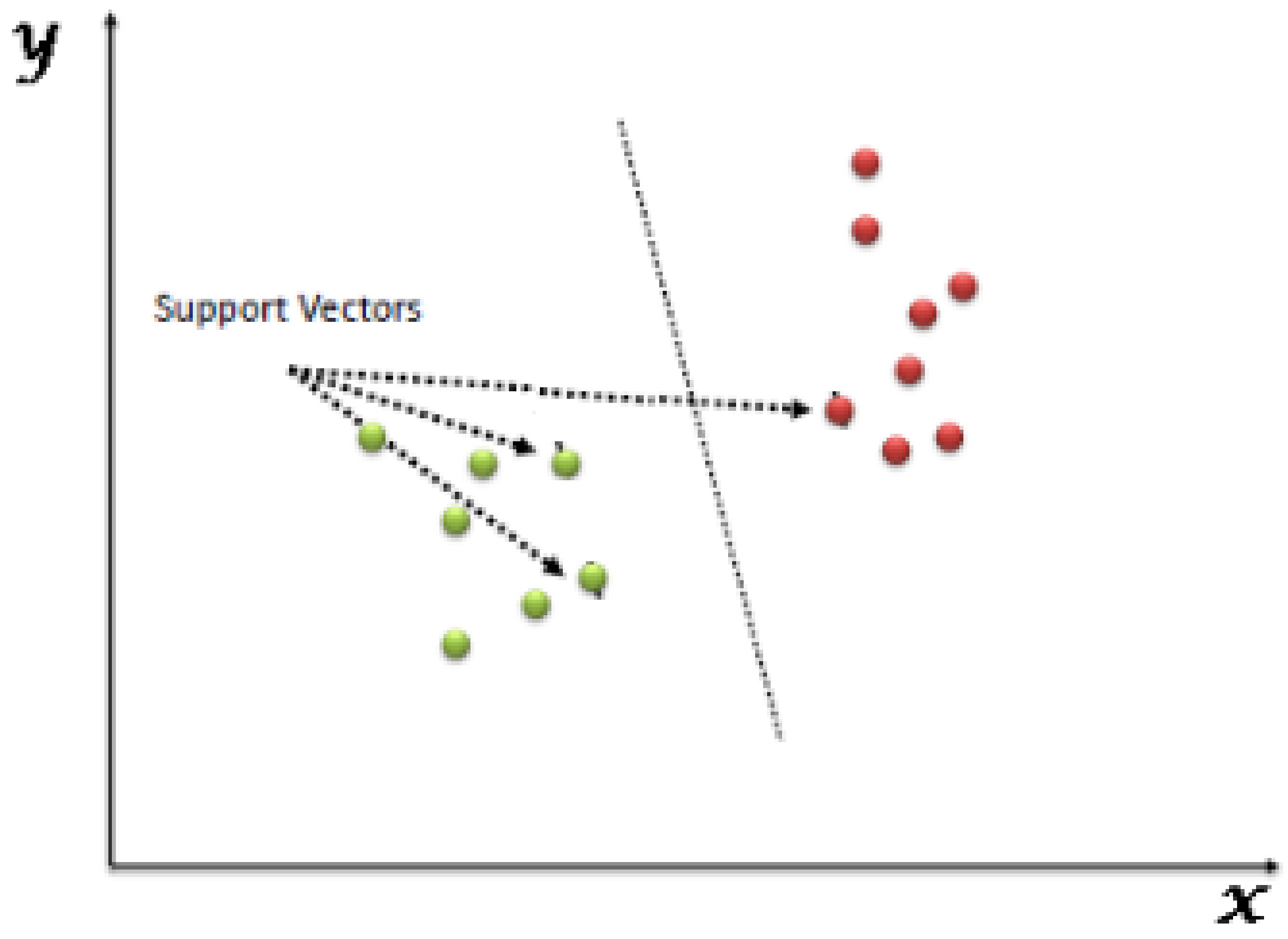
1. SPLITTING WORD2VEC FEATURES:
2. TRAINING THE LOGISTIC REGRESSION MODEL
3. PREDICTING ON THE VALIDATION SET
4. CALCULATING THE F1 SCORE FOR THE VALIDATION SET



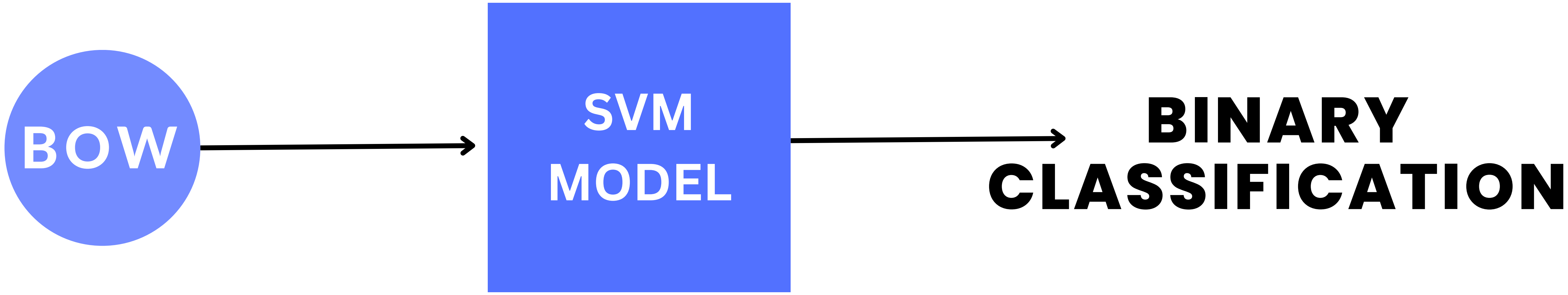
USING LOGISTIC REGRESSION AND W2V WE OBTAINED ACCURACY OF 0.61

1. SUPPORT VECTOR MACHINE

THE ALGORITHM IS A CLASSIFICATION TECHNIQUE THAT OPERATES IN AN N-DIMENSIONAL SPACE, WHERE N REPRESENTS THE NUMBER OF FEATURES AVAILABLE FOR EACH DATA ITEM. THE ALGORITHM AIMS TO SEPARATE TWO CLASSES BY IDENTIFYING A HYPERPLANE THAT DISTINGUISHES THEM.



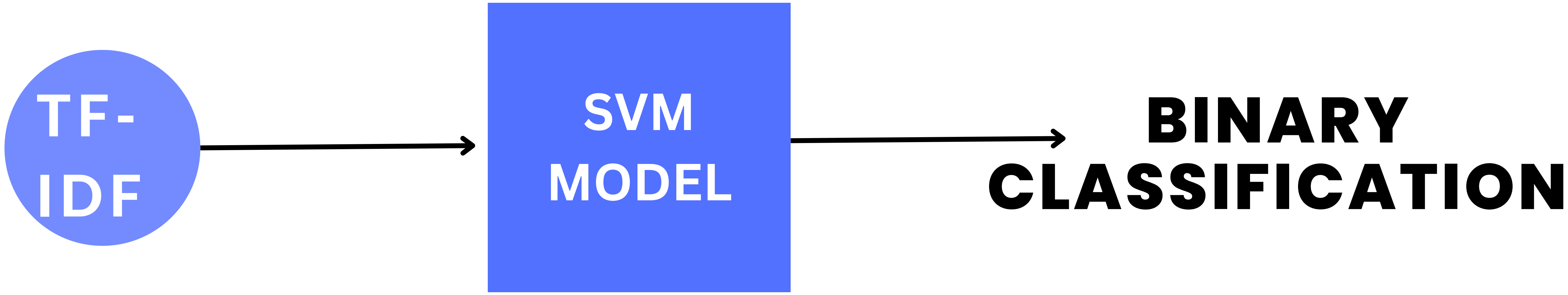
1. SVM-BOW




```
  tweets_df = train.append(test, ignore_index=True, sort=True)
/Users/mohamed/Documents/NLP_MS1/code_1.py:56: FutureWarning: The default value of regex will change
  tweets_df['clean_tweet'] = tweets_df.clean_tweet.str.replace("[^a-zA-Z#]", " ")
0.5088207985143919
```

USING SUPPORT VECTOR MACHINE AND BOW WE OBTAINED ACCURACY OF 0.5081

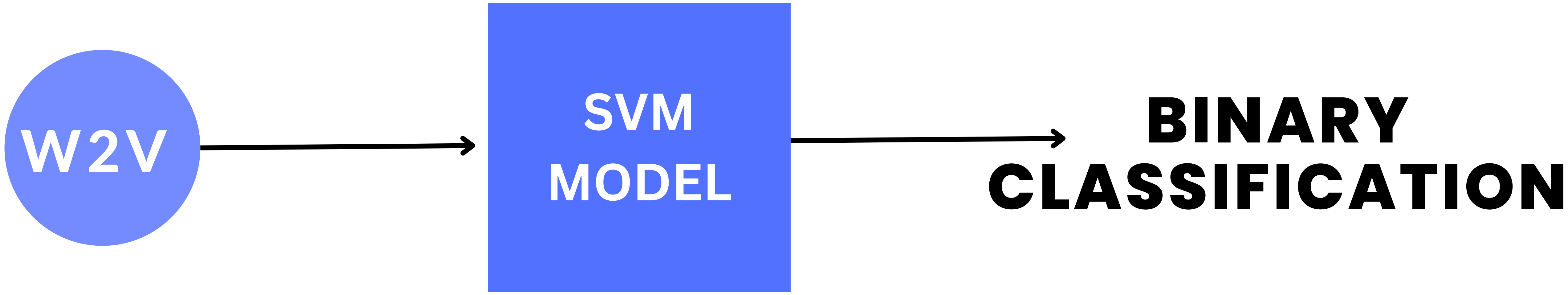
1. SVM-TFIDF




```
/Users/mohamed/Documents/NLP_MS1/code_1.py:56: FutureWarning: The default value of regex will
  tweets_df['clean_tweet'] = tweets_df.clean_tweet.str.replace("[^a-zA-Z#]", " ")
0.5127272727272728
```

USING SUPPORT VECTOR MACHINE AND TF-IDF WE OBTAINED ACCURACY OF 0.512

1. SVM-W2V

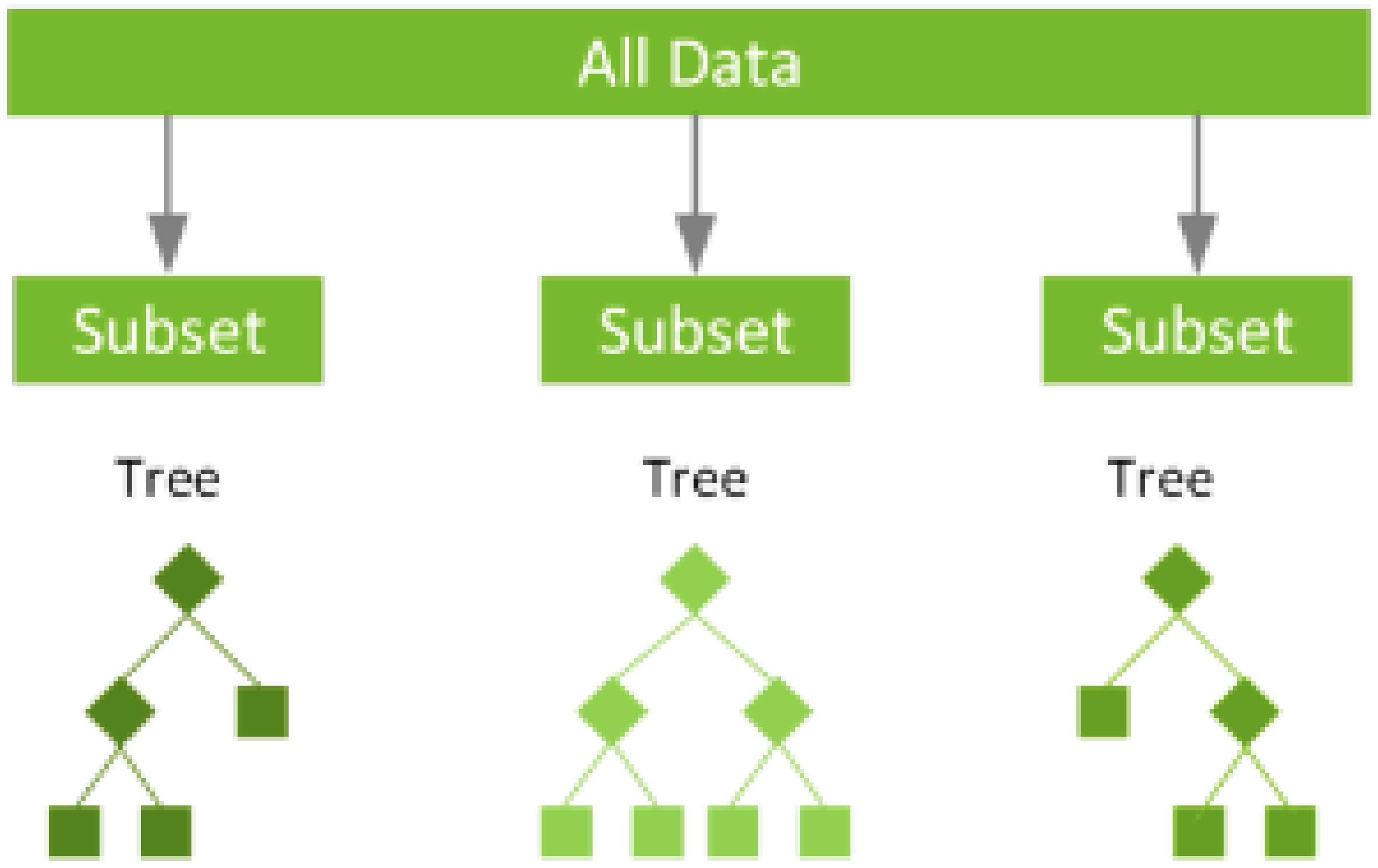



```
/Users/mohamed/Documents/NLP_MS1/code_1.py:56: FutureWarning: The default value of regex w
  tweets_df['clean_tweet'] = tweets_df.clean_tweet.str.replace("[^a-zA-Z#]", " ")
0.6146645865834633
```

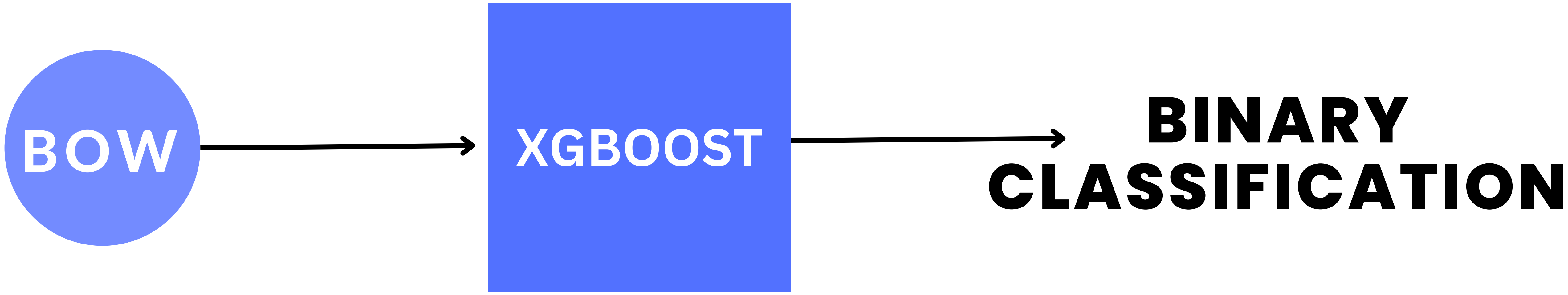
USING SUPPORT VECTOR MACHINE AND W2V WE OBTAINED ACCURACY OF 0.615

1. XGBOOST

IN THIS MODEL WE BUILD DIFFEERENT DECSISION TREES WHICH ACTS AS DIFFERENT MODELS IN PARALELL AND THE DIFFERENCE IN THESE TREES ARE HOW THEY ARE BUILT FROM THE FEATURES GIVEN AND BASED ON EVERY DESCISION TREE PREDICION



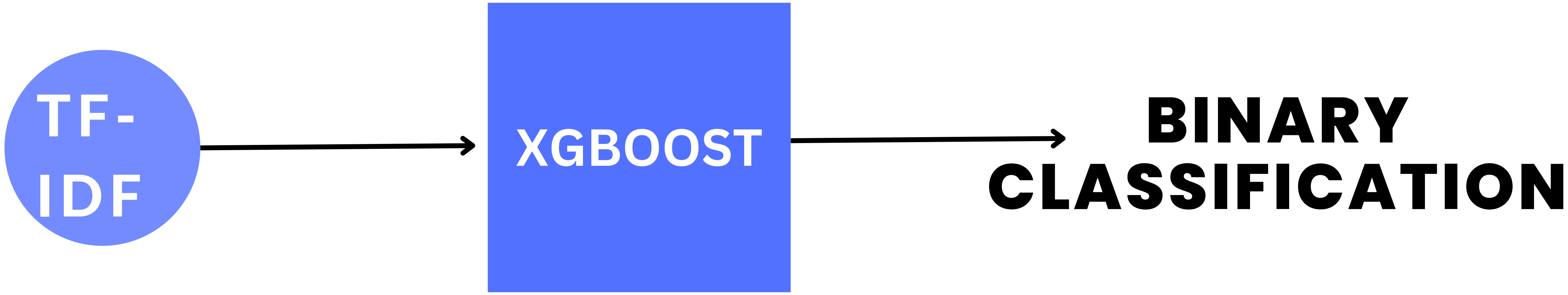
1. XGBOOST-BOW




```
  tweets_df['clean_tweet'] = tweets_df.clean_tweet.str.replace("[^a-zA-Z#]", " ")
0.5247706422018349
```

USING XGBOOST AND BOW WE OBTAINED ACCURACY OF 0.52

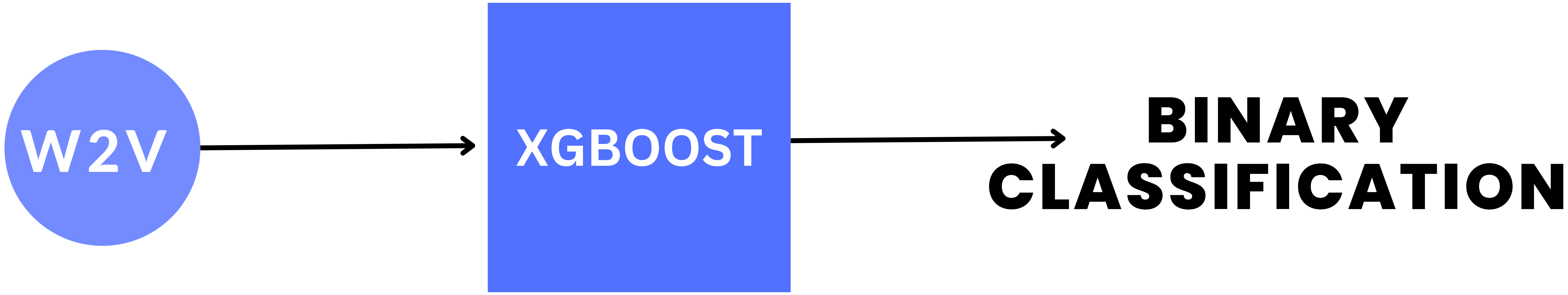
1. XGBOOST-TFIDF



```
  tweets_df['clean_tweet'] = tweets_df.clean_tweet.str.replace("[^a-zA-Z#]", " ")
0.5394265232974911
```

USING XGBOOST AND TF-IDF WE OBTAINED ACCURACY OF 0.54

1. XGBOOST-W2V




```
  tweets_df['clean_tweet'] = tweets_df.clean_tweet.str.replace("[^a-zA-Z#]", " ")
0.6522911051212937
```

USING XGBOOST AND W2V WE OBTAINED ACCURACY OF 0.65

1. RESULTS

XGBOOST-W2V

BASED ON OUR FINDINGS USING WORD2VEC WITH XGBOOST MODEL GAVE US THE MOST ACCURATE PREDICTIONS WITH F SCORE = 65.2 AND THIS ACCURACY CAN BE IMPROVED USING BIGGER DATASETS IN ORDER TO CAPTURE MORE SEMANTIC AND SYNTATIC MEANING OF EACH WORD AND BE ABLE TO DIFFRENTIATE BETWEEN THE TWO CLASSES MORE ACCURATE

THANK YOU

Mohamed Ashraf
46-0831