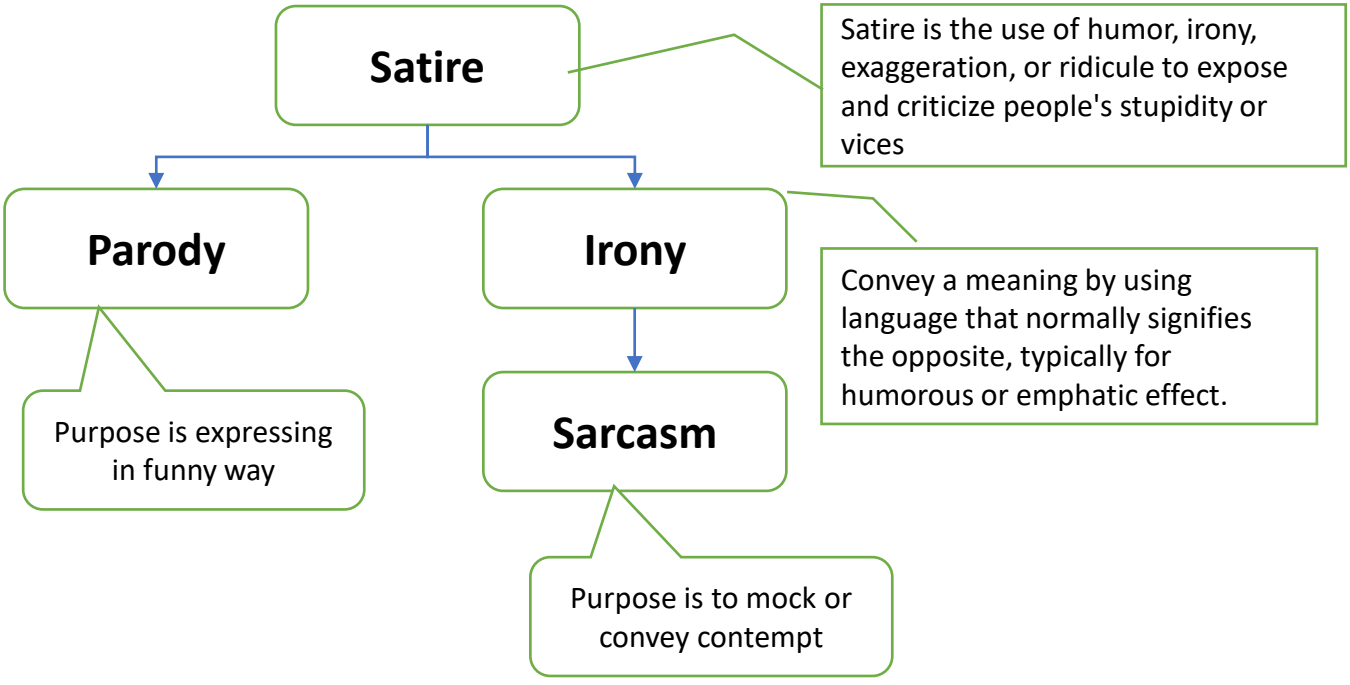
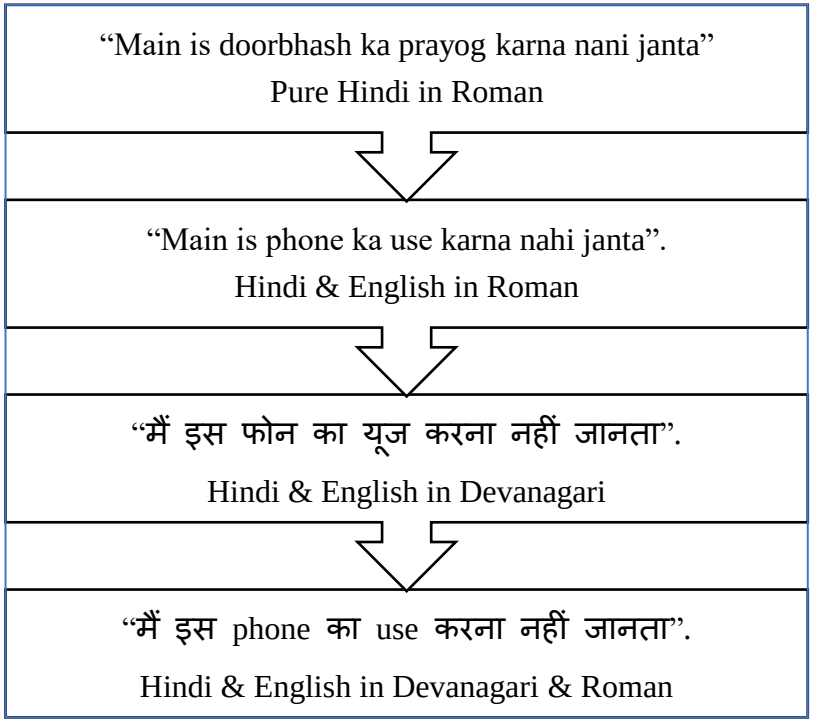
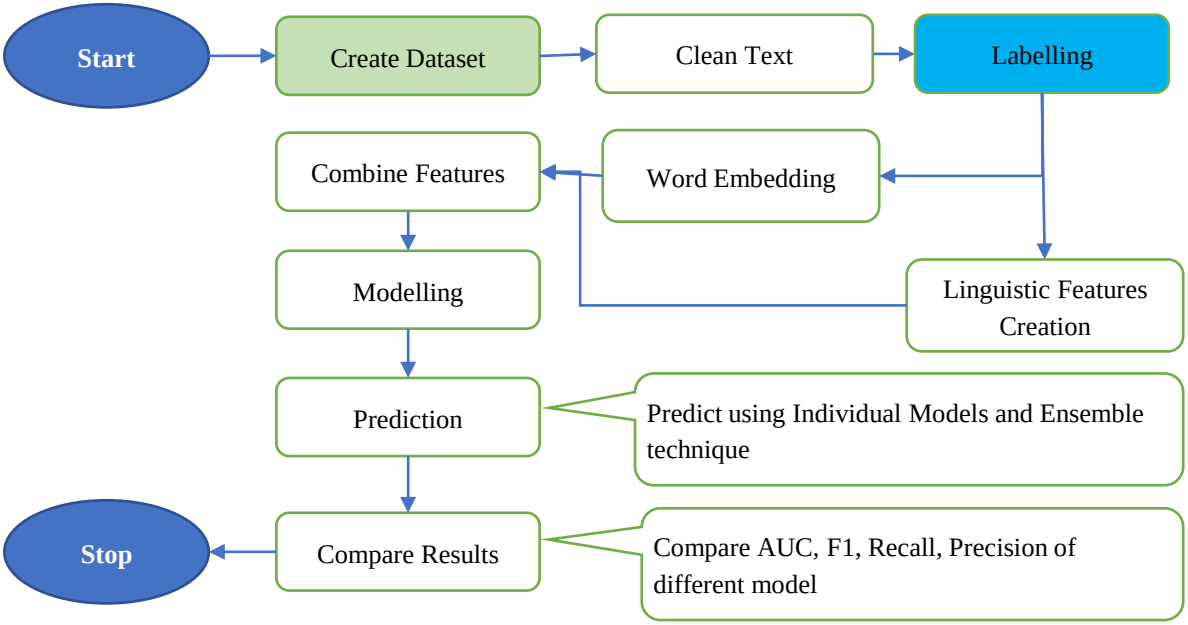
Relationship between Sarcasm & Satire



Evolution of Hinglish from Hindi



Overall Approach - SDSHL



Classifiers

1. Logistic Regression
2. Light Gradient Boost Model
3. Naïve Bayesian
4. Support Vector Machine
5. AdaBoost Classifier
6. Gradient Boost Classifier
7. Random Forest Classifier
8. Perceptron (Neural Network)

Word Embedding

1. TFIDF
2. Word2Vec
3. BOW
4. IndicBERT
5. Multilingual BERT
6. fastText
7. fastText Wiki
8. fastText Indicnlp/ IndicFT

Feature Engineering

1. Lexical Feature
2. Combined = IndicFT + LexicalFeature

Text Cleaning Steps

Remove extra space, RT, http / www links, “@” account text information, line break

Remove non Hindi sentences

Remove over expression (extra emoticons),

Remove sentences which are less than 4 words.

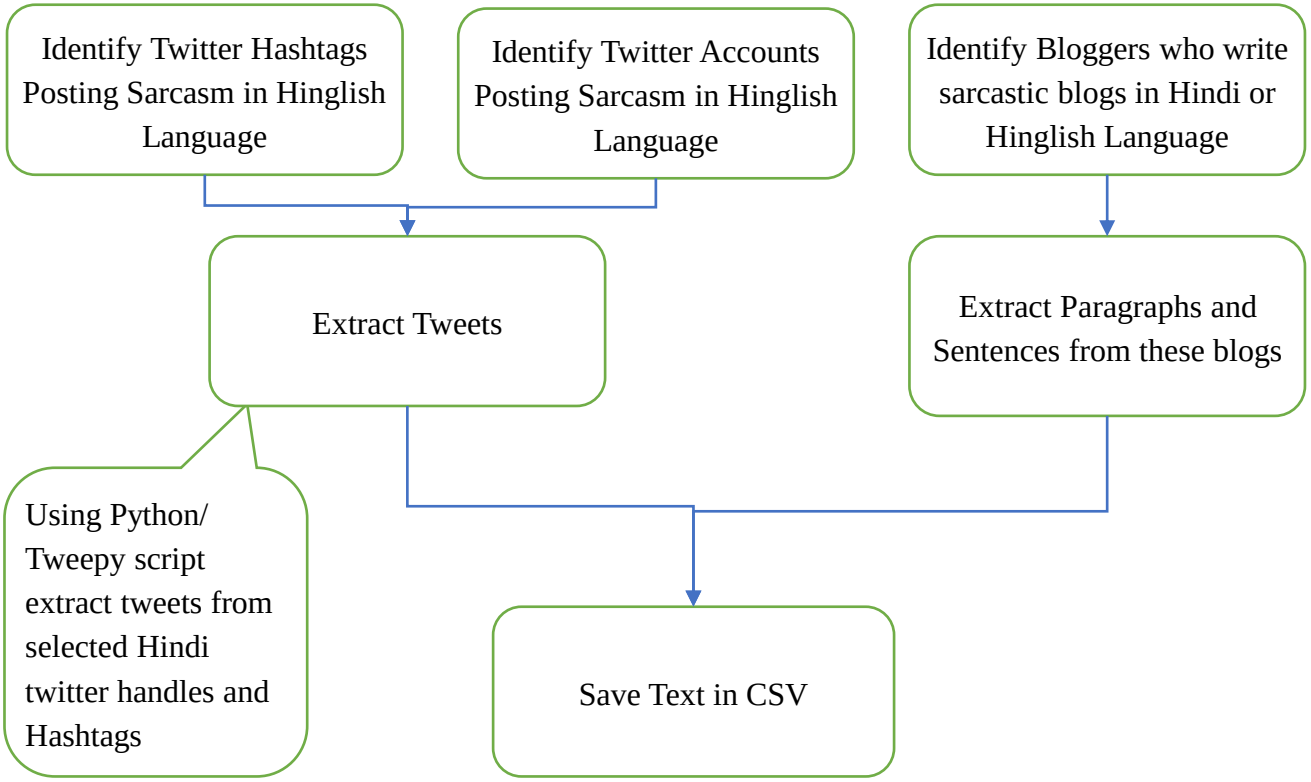
Remove greeting related msg like #सुप्रभात, #राम राम जी, जय श्रीराम etc.

Remove duplicate sentences

Remove national or religious flag information like ININ

Replace Hindi । (Hindi sentence marker with “.”)

Dataset Creation Steps



Sentence Labelling Steps

Share CSV file with 3 Hindi Language expert

Expert Labelling for each Sentecne

Final Labelling based on Maximm Voting

Classification Type - Feature Type

Discussed in Section Number

		Feature Types		
		LFS	Embedding	Both
Classification Type	Rule Based	2.5.1	x	x
	Classical ML Algorithms	2.5.4	2.5.3	2.5.2
	CNN	2.5.9	2.5.6	2.5.5
	Transformers	x	2.5.7	x
	Transfer Learning	x	2.5.8	x

Metrics in ML Project

Model1

Actual	Observation: FALSE	Observation: TRUE	
FALSE	820	30	850
TRUE	70	80	150
	890	110	1000

Accuracy	0.90
Recall	0.73
Precesion	0.53
F1 Score	0.81
Error Rate	0.10

Model2

Actual	Observation: FALSE	Observation: TRUE	
FALSE	805	45	850
TRUE	55	95	150
	860	140	1000

Accuracy	0.90
Recall	0.68
Precesion	0.63
F1 Score	0.77
Error Rate	0.10

Performance based Metrics

Top 10 Best Models

Classifier	Embedding Name	Acc	Recall	Precision	F1	AUC
NB	fasttext_Wiki	0.74	0.76	0.74	0.75	0.81
NB	fasttext_Indicnlp	0.73	0.75	0.72	0.74	0.81
GBC	fasttext_Indicnlp	0.76	0.68	0.81	0.74	0.82
LGBM	fasttext_Indicnlp	0.75	0.67	0.80	0.73	0.81
GBC	combined	0.75	0.67	0.80	0.73	0.81
SVC	fasttext_Indicnlp	0.74	0.68	0.77	0.72	0.81
Perceptron	fasttext_Indicnlp	0.72	0.72	0.71	0.72	0.71
LR	fasttext_Wiki	0.74	0.68	0.76	0.72	0.80
RFC	combined	0.74	0.64	0.81	0.72	0.81
Perceptron	combined	0.61	0.99	0.56	0.72	0.61

Bottom 10 Worse Models

Classifier	Embedding Name	Acc	Recall	Precision	F1	AUC
LR	tfidf	0.57	0.47	0.59	0.53	0.58
LGBM	tfidf	0.55	0.51	0.55	0.53	0.59
GBC	tfidf	0.55	0.50	0.55	0.52	0.58
LR	bow	0.56	0.46	0.58	0.51	0.58
NB	tfidf	0.55	0.46	0.55	0.50	0.57
ADB	bow	0.52	0.47	0.53	0.50	0.56
NB	bow	0.57	0.41	0.60	0.49	0.58
ADB	Tfidf	0.50	0.46	0.51	0.48	0.54
NB	Lexical Features	0.60	0.36	0.71	0.48	0.71
Perceptron	Lexical Features	0.60	0.27	0.79	0.40	0.60

Best Embedding – Average All Metrics of All Classifier

Embedding	Avg Acc	Avg Recall	Avg Precision	Avg F1	Avg AUC
fasttext_Indicnlp	0.73	0.67	0.76	0.71	0.79
combined	0.71	0.71	0.73	0.71	0.77
fasttext_Wiki	0.71	0.65	0.74	0.69	0.78
fasttext	0.61	0.81	0.59	0.67	0.66
BERT_Multilingual	0.64	0.70	0.64	0.66	0.69
word2vec	0.57	0.77	0.56	0.64	0.64
Kera Tokens	0.65	0.63	0.66	0.64	0.69
BERT-Indicnlp	0.62	0.63	0.63	0.62	0.65
BERT_Indicnlp	0.57	0.68	0.56	0.62	0.62
Lexical Features	0.66	0.59	0.69	0.62	0.71
tfidf	0.56	0.57	0.56	0.55	0.58
bow	0.55	0.54	0.57	0.54	0.58

Best Classifier – Average All Metrics of All Embedding

Classifier	Avg Acc	Avg Recall	Avg Precision	Avg F1	Avg AUC
BERT_Pytorch_TT2	0.74	0.69	0.76	0.72	0.80
fastText_wiki_TT	0.72	0.66	0.75	0.70	0.79
fastText_indicnlp_TT	0.70	0.63	0.73	0.68	0.79
RFC	0.66	0.67	0.66	0.66	0.72
CNN	0.68	0.62	0.70	0.66	0.73
GBC	0.64	0.67	0.64	0.65	0.68
SVC	0.64	0.66	0.64	0.65	0.70
LGBM	0.64	0.64	0.65	0.64	0.69
LR	0.64	0.65	0.65	0.64	0.68
RNN	0.62	0.63	0.62	0.62	0.65
BERT_Transformer_TT2	0.58	0.68	0.57	0.62	0.62
ADB	0.61	0.62	0.62	0.62	0.66
NB	0.59	0.68	0.61	0.61	0.68
Perceptron	0.59	0.69	0.61	0.61	0.58

Task Transfer Learning

Embedding Name	Acc	Recall	Precision	F1	AUC
BERT_Multilingual (Pytorch)	0.74	0.69	0.76	0.72	0.8
BERT_Multilingual (Transformer)	0.58	0.67	0.57	0.62	0.61
BERT_Indicnlp (Transformer)	0.57	0.68	0.56	0.62	0.62
fasttext_Indicnlp	0.7	0.63	0.73	0.68	0.79
fasttext_Wiki	0.72	0.66	0.75	0.7	0.79

Embedding Based Metrics

Embedding Transfer: BERT Multilingual

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.71	0.71	0.71	0.71	0.73
LGBM	0.62	0.61	0.62	0.62	0.71
NB	0.56	0.80	0.54	0.64	0.66
SVC	0.67	0.65	0.68	0.66	0.74
ADB	0.62	0.62	0.61	0.62	0.68
GBC	0.64	0.63	0.65	0.64	0.66
RFC	0.66	0.64	0.67	0.66	0.72
Perceptron	0.58	0.97	0.54	0.70	0.58

Embedding Transfer: BERT Indicnlp

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.60	0.64	0.59	0.61	0.61
LGBM	0.68	0.67	0.68	0.68	0.70
NB	0.58	0.66	0.57	0.61	0.65
SVC	0.60	0.67	0.58	0.62	0.62
ADB	0.59	0.58	0.59	0.59	0.61
GBC	0.64	0.67	0.64	0.65	0.68
RFC	0.68	0.69	0.68	0.69	0.71
Perceptron	0.62	0.44	0.69	0.54	0.62

Embedding Transfer: fastText Indicnlp

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.73	0.66	0.77	0.71	0.81
LGBM	0.75	0.67	0.80	0.73	0.81
NB	0.73	0.75	0.72	0.74	0.81
SVC	0.74	0.68	0.77	0.72	0.81
ADB	0.70	0.64	0.74	0.68	0.78
GBC	0.76	0.68	0.81	0.74	0.82
RFC	0.73	0.62	0.79	0.70	0.80
Perceptron	0.72	0.72	0.71	0.72	0.71

Embedding Transfer: fastText Wiki

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.74	0.68	0.76	0.72	0.80
LGBM	0.71	0.65	0.74	0.69	0.78
NB	0.74	0.76	0.74	0.75	0.81
SVC	0.72	0.65	0.76	0.70	0.80
ADB	0.70	0.63	0.73	0.68	0.77
GBC	0.72	0.65	0.76	0.70	0.79
RFC	0.71	0.64	0.74	0.69	0.78
Perceptron	0.66	0.57	0.69	0.62	0.66

fastText Embedding

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.60	0.84	0.57	0.68	0.64
LGBM	0.65	0.77	0.62	0.69	0.68
NB	0.51	0.97	0.51	0.66	0.66
SVC	0.59	0.77	0.57	0.65	0.67
ADB	0.62	0.70	0.61	0.65	0.66
GBC	0.66	0.77	0.63	0.69	0.70
RFC	0.66	0.75	0.64	0.69	0.71
Perceptron	0.55	0.88	0.53	0.66	0.54

Word2Vec Embedding

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.58	0.79	0.56	0.65	0.63
LGBM	0.58	0.71	0.56	0.63	0.66
NB	0.51	0.96	0.51	0.66	0.65
SVC	0.56	0.72	0.55	0.62	0.66
ADB	0.60	0.82	0.57	0.67	0.63
GBC	0.58	0.81	0.55	0.66	0.59
RFC	0.60	0.73	0.57	0.64	0.68
Perceptron	0.58	0.64	0.58	0.61	0.59

BOW Embedding

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.56	0.46	0.58	0.51	0.58
LGBM	0.58	0.53	0.59	0.56	0.60
NB	0.57	0.41	0.60	0.49	0.58
SVC	0.59	0.53	0.60	0.56	0.63
ADB	0.52	0.47	0.53	0.50	0.56
GBC	0.54	0.55	0.54	0.54	0.60
RFC	0.61	0.61	0.61	0.61	0.64
Perceptron	0.46	0.73	0.47	0.57	0.45

TFIDF Embedding

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.57	0.47	0.59	0.53	0.58
LGBM	0.55	0.51	0.55	0.53	0.59
NB	0.55	0.46	0.55	0.50	0.57
SVC	0.60	0.58	0.60	0.59	0.63
ADB	0.50	0.46	0.51	0.48	0.54
GBC	0.55	0.50	0.55	0.52	0.58
RFC	0.62	0.61	0.62	0.61	0.67
Perceptron	0.50	1.00	0.50	0.67	0.50

Lexical Feature Engineering

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.68	0.57	0.72	0.64	0.76
LGBM	0.66	0.66	0.66	0.66	0.70
NB	0.60	0.36	0.71	0.48	0.71
SVC	0.68	0.72	0.67	0.69	0.73
ADB	0.68	0.69	0.67	0.68	0.73
GBC	0.66	0.73	0.64	0.68	0.73
RFC	0.68	0.72	0.66	0.69	0.74
Perceptron	0.60	0.27	0.79	0.40	0.60

Combined Embedding: fastText Indicnlp + Lexical Features

Classifier	Acc	Recall	Precision	F1	AUC
LR	0.72	0.66	0.75	0.7	0.82
LGBM	0.74	0.63	0.81	0.71	0.81
NB	0.71	0.72	0.71	0.71	0.8
SVC	0.7	0.73	0.69	0.71	0.75
ADB	0.69	0.62	0.72	0.67	0.78
GBC	0.75	0.67	0.8	0.73	0.81
RFC	0.74	0.64	0.81	0.72	0.81
Perceptron	0.61	0.99	0.56	0.72	0.61

Transfer Learning Based Metrics

Top 10 – No Transfer Learning Models

Classifier	Embedding Name	Acc	Recall	Precision	F1	AUC
LGBM	fasttext	0.65	0.77	0.62	0.69	0.68
GBC	fasttext	0.66	0.77	0.63	0.69	0.70
RFC	fasttext	0.66	0.75	0.64	0.69	0.71
SVC	Lexical Features	0.68	0.72	0.67	0.69	0.73
RFC	Lexical Features	0.68	0.72	0.66	0.69	0.74
LR	fasttext	0.60	0.84	0.57	0.68	0.64
ADB	Lexical Features	0.68	0.69	0.67	0.68	0.73
GBC	Lexical Features	0.66	0.73	0.64	0.68	0.73
Perceptron	tfidf	0.50	1.00	0.50	0.67	0.50
ADB	word2vec	0.60	0.82	0.57	0.67	0.63

Top 10 – Transfer Learning Models

Classifier	TL Type	Embedding Name	Acc	Recall	Precision	F1	AUC
NB	EMB	fasttext_Wiki	0.74	0.76	0.74	0.75	0.81
NB	EMB	fasttext_Indicnlp	0.73	0.75	0.72	0.74	0.81
GBC	EMB	fasttext_Indicnlp	0.76	0.68	0.81	0.74	0.82
LGBM	EMB	fasttext_Indicnlp	0.75	0.67	0.80	0.73	0.81
GBC	EMB	combined	0.75	0.67	0.80	0.73	0.81
SVC	EMB	fasttext_Indicnlp	0.74	0.68	0.77	0.72	0.81
Perceptron	EMB	fasttext_Indicnlp	0.72	0.72	0.71	0.72	0.71
LR	EMB	fasttext_Wiki	0.74	0.68	0.76	0.72	0.80
RFC	EMB	combined	0.74	0.64	0.81	0.72	0.81
Perceptron	EMB	combined	0.61	0.99	0.56	0.72	0.61

Classifier Based Metrics

Logistic Regression

Embedding Name	Acc	Recall	Precision	F1	AUC
fasttext_Wiki	0.74	0.68	0.76	0.72	0.80
BERT_Multilingual	0.71	0.71	0.71	0.71	0.73
fasttext_Indicnlp	0.73	0.66	0.77	0.71	0.81
combined	0.72	0.66	0.75	0.70	0.82
fasttext	0.60	0.84	0.57	0.68	0.64
word2vec	0.58	0.79	0.56	0.65	0.63
Lexical Features	0.68	0.57	0.72	0.64	0.76
BERT-Indicnlp	0.60	0.64	0.59	0.61	0.61
tfidf	0.57	0.47	0.59	0.53	0.58
bow	0.56	0.46	0.58	0.51	0.58

Naïve Bayesian

Embedding Name	Acc	Recall	Precision	F1	AUC
fasttext_Wiki	0.74	0.76	0.74	0.75	0.81
fasttext_Indicnlp	0.73	0.75	0.72	0.74	0.81
combined	0.71	0.72	0.71	0.71	0.80
word2vec	0.51	0.96	0.51	0.66	0.65
fasttext	0.51	0.97	0.51	0.66	0.66
BERT_Multilingual	0.56	0.80	0.54	0.64	0.66
BERT-Indicnlp	0.58	0.66	0.57	0.61	0.65
tfidf	0.55	0.46	0.55	0.50	0.57
bow	0.57	0.41	0.60	0.49	0.58
Lexical Features	0.60	0.36	0.71	0.48	0.71

Gradient Boost Classifier (GBC)

Embedding Name	Acc	Recall	Precision	F1	AUC
fasttext_Indicnlp	0.76	0.68	0.81	0.74	0.82
combined	0.75	0.67	0.80	0.73	0.81
fasttext_Wiki	0.72	0.65	0.76	0.70	0.79
fasttext	0.66	0.77	0.63	0.69	0.70
Lexical Features	0.66	0.73	0.64	0.68	0.73
word2vec	0.58	0.81	0.55	0.66	0.59
BERT-Indicnlp	0.64	0.67	0.64	0.65	0.68
BERT_Multilingual	0.64	0.63	0.65	0.64	0.66
bow	0.54	0.55	0.54	0.54	0.60
tfidf	0.55	0.50	0.55	0.52	0.58

Light Gradient Boost Model (LGBM)

Embedding Name	Acc	Recall	Precision	F1	AUC
fasttext_Indicnlp	0.75	0.67	0.80	0.73	0.81
combined	0.74	0.63	0.81	0.71	0.81
fasttext	0.65	0.77	0.62	0.69	0.68
fasttext_Wiki	0.71	0.65	0.74	0.69	0.78
BERT-Indicnlp	0.68	0.67	0.68	0.68	0.70
Lexical Features	0.66	0.66	0.66	0.66	0.70
word2vec	0.58	0.71	0.56	0.63	0.66
BERT_Multilingual	0.62	0.61	0.62	0.62	0.71
bow	0.58	0.53	0.59	0.56	0.60
tfidf	0.55	0.51	0.55	0.53	0.59

Random Forest Classifier (RFC)

Embedding Name	Acc	Recall	Precision	F1	AUC
combined	0.74	0.64	0.81	0.72	0.81
fasttext_Indicnlp	0.73	0.62	0.79	0.70	0.80
BERT-Indicnlp	0.68	0.69	0.68	0.69	0.71
fasttext	0.66	0.75	0.64	0.69	0.71
fasttext_Wiki	0.71	0.64	0.74	0.69	0.78
Lexical Features	0.68	0.72	0.66	0.69	0.74
BERT_Multilingual	0.66	0.64	0.67	0.66	0.72
word2vec	0.60	0.73	0.57	0.64	0.68
tfidf	0.62	0.61	0.62	0.61	0.67
bow	0.61	0.61	0.61	0.61	0.64

Support Vector Machine (SVM)

Embedding Name	Acc	Recall	Precision	F1	AUC
fasttext_Indicnlp	0.74	0.68	0.77	0.72	0.81
combined	0.70	0.73	0.69	0.71	0.75
fasttext_Wiki	0.72	0.65	0.76	0.70	0.80
Lexical Features	0.68	0.72	0.67	0.69	0.73
BERT_Multilingual	0.67	0.65	0.68	0.66	0.74
fasttext	0.59	0.77	0.57	0.65	0.67
word2vec	0.56	0.72	0.55	0.62	0.66
BERT-Indicnlp	0.60	0.67	0.58	0.62	0.62
tfidf	0.60	0.58	0.60	0.59	0.63
bow	0.59	0.53	0.60	0.56	0.63

Perceptron

Embedding Name	Acc	Recall	Precision	F1	AUC
fasttext_Indicnlp	0.72	0.72	0.71	0.72	0.71
combined	0.61	0.99	0.56	0.72	0.61
BERT_Multilingual	0.58	0.97	0.54	0.70	0.58
tfidf	0.50	1.00	0.50	0.67	0.50
fasttext	0.55	0.88	0.53	0.66	0.54
fasttext_Wiki	0.66	0.57	0.69	0.62	0.66
word2vec	0.58	0.64	0.58	0.61	0.59
bow	0.46	0.73	0.47	0.57	0.45
BERT-Indicnlp	0.62	0.44	0.69	0.54	0.62
Lexical Features	0.60	0.27	0.79	0.40	0.60

AdaBoost

Embedding Name	Acc	Recall	Precision	F1	AUC
fasttext_Indicnlp	0.70	0.64	0.74	0.68	0.78
fasttext_Wiki	0.70	0.63	0.73	0.68	0.77
Lexical Features	0.68	0.69	0.67	0.68	0.73
word2vec	0.60	0.82	0.57	0.67	0.63
combined	0.69	0.62	0.72	0.67	0.78
fasttext	0.62	0.70	0.61	0.65	0.66
BERT_Multilingual	0.62	0.62	0.61	0.62	0.68
BERT-Indicnlp	0.59	0.58	0.59	0.59	0.61
bow	0.52	0.47	0.53	0.50	0.56
tfidf	0.50	0.46	0.51	0.48	0.54

CNN and RNN

Classifier	Acc	Recall	Precision	F1	AUC
CNN	0.68	0.62	0.7	0.66	0.73
RNN	0.62	0.63	0.62	0.62	0.65