Tweet Classification

MURALI KRISHNA VALLURI (671441135)

SPOORTHI PENDYALA (679836791)

Pre - Processing Steps

- Initial excel sheet is converted to csv file by removing rows with class values ‘2’, ‘IR’, ‘Irrelevant’, ‘!!!!’. Also Date and Time columns where removed.
- Tweet is converted to lowercase.
- Removed special characters and replaced ‘#’, ‘@’, ‘””, ‘’’ with blank characters, hyperlinks and html tags.
- Stripped the text which removes extra spaces at the beginning and end of the text and finally we removed the stop words from the text.

What has changed?

- Changed the encoding type of input csv file to UTF-8 which preserved all tweets.
- Performed stemming using NLTK library.
- Replaced emoticons, slang text using predefined set of values obtained from online sources.

Classification Algorithms

- Decision Trees
- Logistic Regression
- Multinomial Naïve Bayes
- K – Nearest
- SVM
- Voting

Note: The results were obtained via 10-fold cross validation on the training data.

Results

- **Obama Data**

Algorithms	Negative Class			Positive Class			Accuracy
	Precision	Recall	F-Score	Precision	Recall	F-Score	
Previous Results	0.57	0.68	0.61	0.64	0.58	0.61	58.47
Logistic Regression	0.57	0.66	0.6	0.63	0.57	0.59	58.53
SVM	0.59	0.64	0.6	0.66	0.61	0.62	58.95
Multinomial NB	0.52	0.77	0.6	0.68	0.49	0.56	55.82
Voting (2,3,1)	**0.59**	**0.67**	**0.61**	**0.63**	**0.59**	**0.61**	**58.98**

- **Romney Data**

Algorithms	Negative Class			Positive Class			Accuracy
	Precision	Recall	F-Score	Precision	Recall	F-Score	
Previous Results	0.59	0.88	0.69	0.64	0.28	0.39	57.21
Logistic Regression	0.58	0.86	0.68	0.65	0.27	0.37	56.08
SVM	0.66	0.61	0.62	0.47	0.54	0.49	54.74
Multinomial NB	0.53	0.98	0.67	0.79	0.07	0.12	53.1
Voting (3,9,1)	**0.60**	**0.83**	**0.68**	**0.60**	**0.37**	**0.45**	**57.05**