INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2020

HATE SPEECH DETECTION AND INSINCERE QUESTION CLASSIFICATION

ABSTRACT

CIQ or Classification of Insincere Question task in FIRE 2019 focuses on differentiating proper information seeking questions from different kinds of insincere questions. **HASOC** shared task attempts for automatic detection of abusive language on Twitter in English, German and Hindi languages. AdaBoost performed best for CIQ task. Our best performing model in HASOC was an ensemble model of SVM, Random Forest and Adaboost classifiers with majority voting.

INTRODUCTION

CIQ - Different toxic, malicious, hate related posts throw the biggest challenges to most community question answering forums.This task attempts to filter out malicious content from the forum of Quora (https://www.quora.com) that will keep their platform more secured for users. The 6 classes include questions related to rhetorical, sexual content,hate speech,hypothetical,others and not insincere content.

HASOC - Social media is a great platform to communicate with people from different demographic groups. People spend considerable amount of time on these forums. Recent studies suggest that most of the online content generated on these platforms contains different forms of abusive language. Task1 is a binary classification for predicting HOF or NOT. Task2 is 4 class classification task between HATE, NONE, OFFN, PRFN and task3 deals with classification for targetted insult as NONE, TIN, UNT.

CORPUS DETAILS

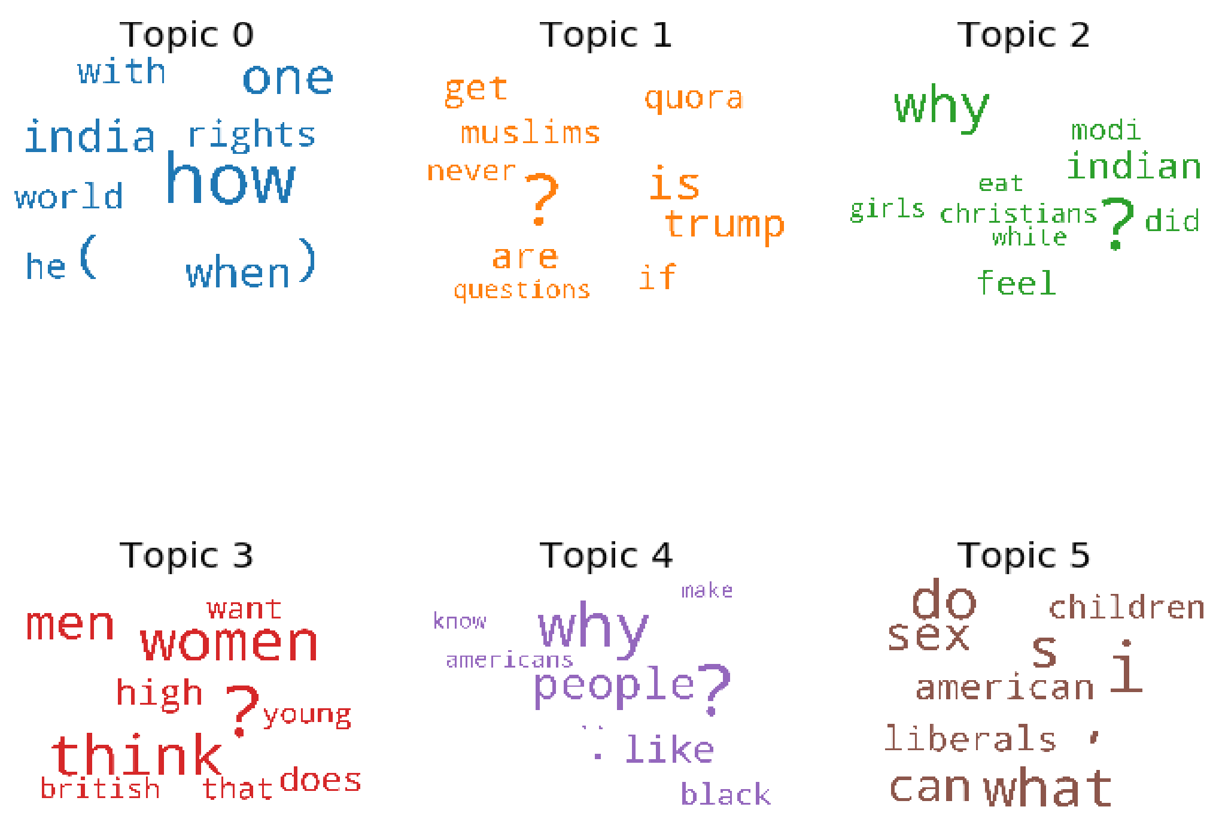
#Samples	#Classes	Tasks
899	6	1

CIQ

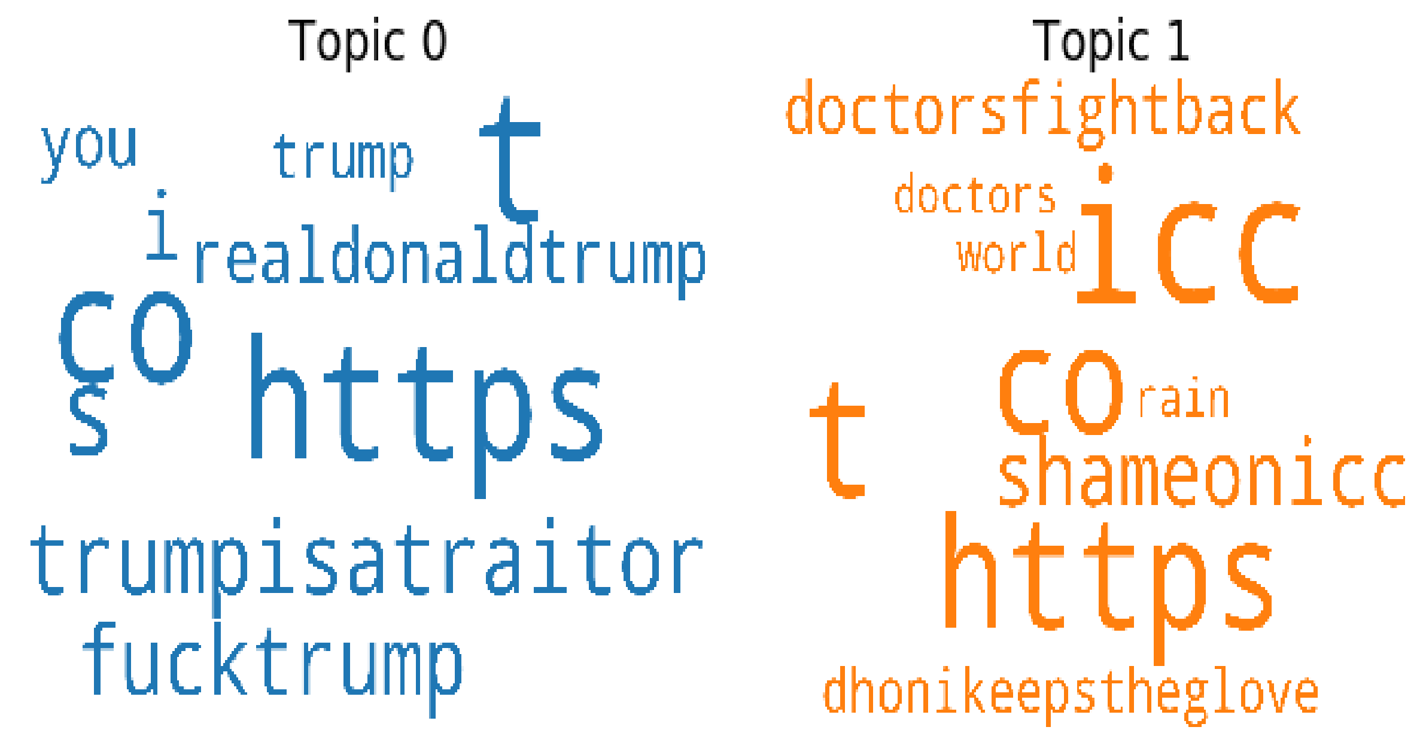
Language	#Samples		#Classes	
		Task1	Task2	Task3
English	5852	2	3	4
German	3819	2	3	-
Hindi	4665	2	3	4

HASOC

WORD CLUSTERS



CIQ



HASOC

RESULTS

Model	Features	Accuracy
GB+3NN+RF	Word uni+bi	62.37
Adaboost		66.33

CIQ

Lang	Task#	Model	Features	F1
EN	1	SVM+RF+AB	Word uni+char2-5+tweetLength	0.77
	2			0.73
	3			0.75
DE	1	SVM+RF+AB	Word uni+char2-5+tweetLength	0.77
	2			0.77
HI	1	SVM+RF+AB	Word uni+char2-5	0.80
	2			0.65
	3			0.74

HASOC

FUTURE WORK

Huge amounts of unlabeled questions from Quora can be explored to improve the clustering techniques and improve the classification. We can explore unsupervised techniques on raw tweets for learning a better representation of implicit form of hate speech. Convolutional neural networks (CNN) could be used to model the interactions between character n-grams in the tweets.

Authors:Vandan Mujadia, Pruthwik Mishra, Dipti Misra Sharma

Research Center Name:MT&NLP Lab, LTRC