INSE - 6180 Project Presentation

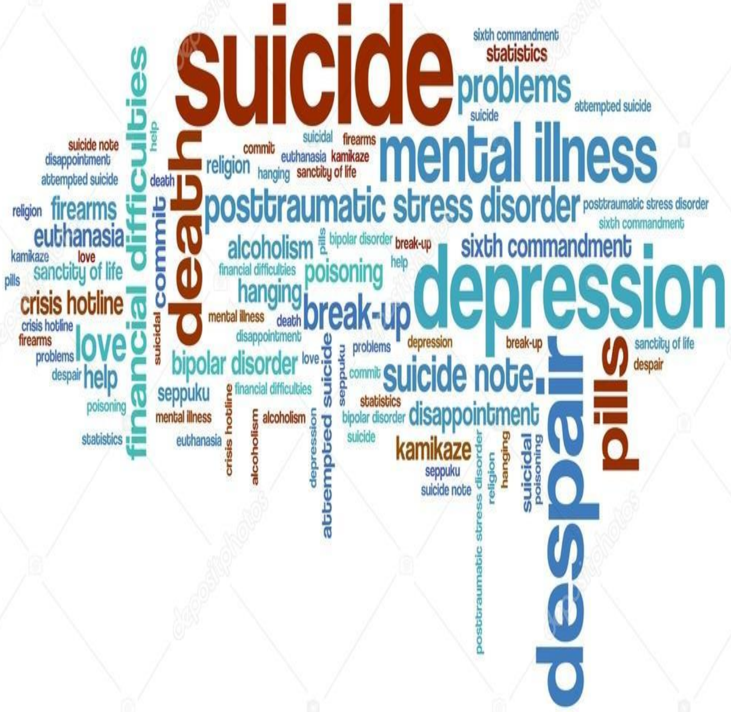
Sentiment Analysis of User Generated Online Content to Detect Suicidal Tendencies

By:

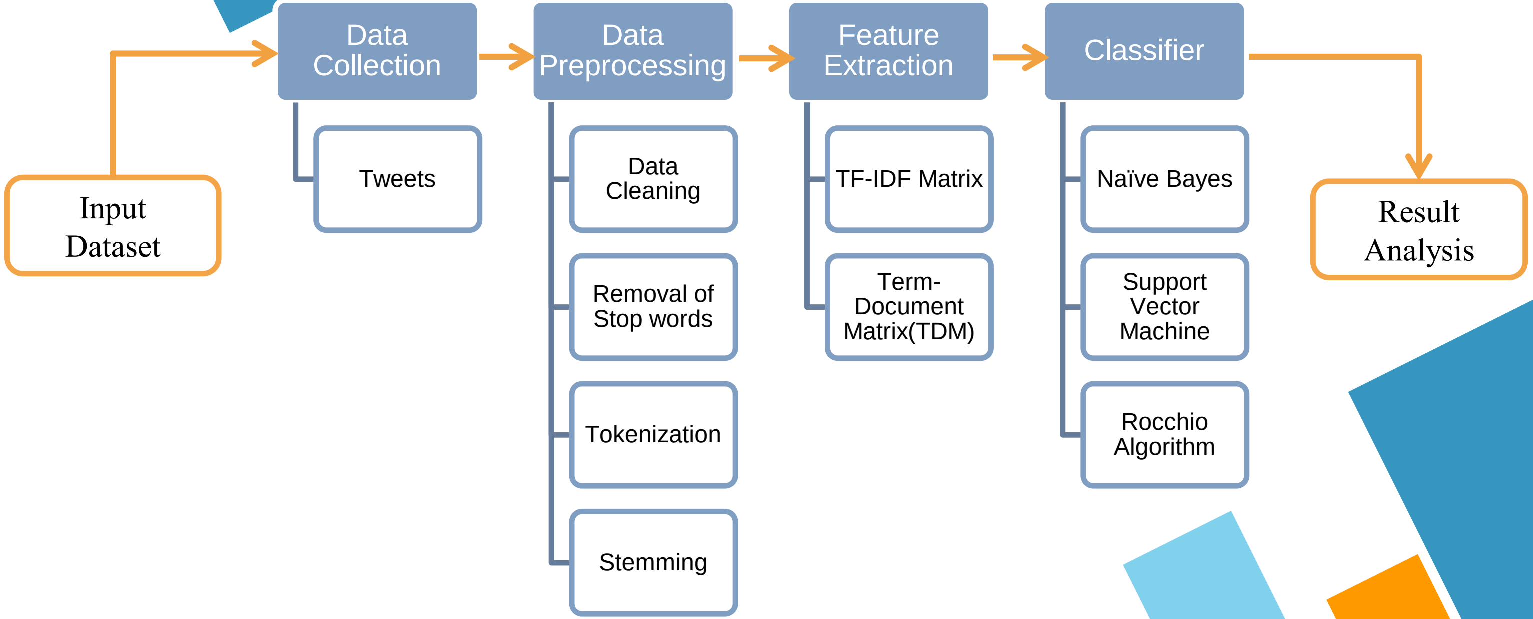
Vasu Jain	Simran Sidhu	Mandeep Kaur
S_Id - 40057063	S_Id - 40011611	S_Id - 40059801
v_ja@encs.concordia.ca	s_idhu@encs.concordia.ca	k_ndeep@encs.concordia.ca

Problem Statement

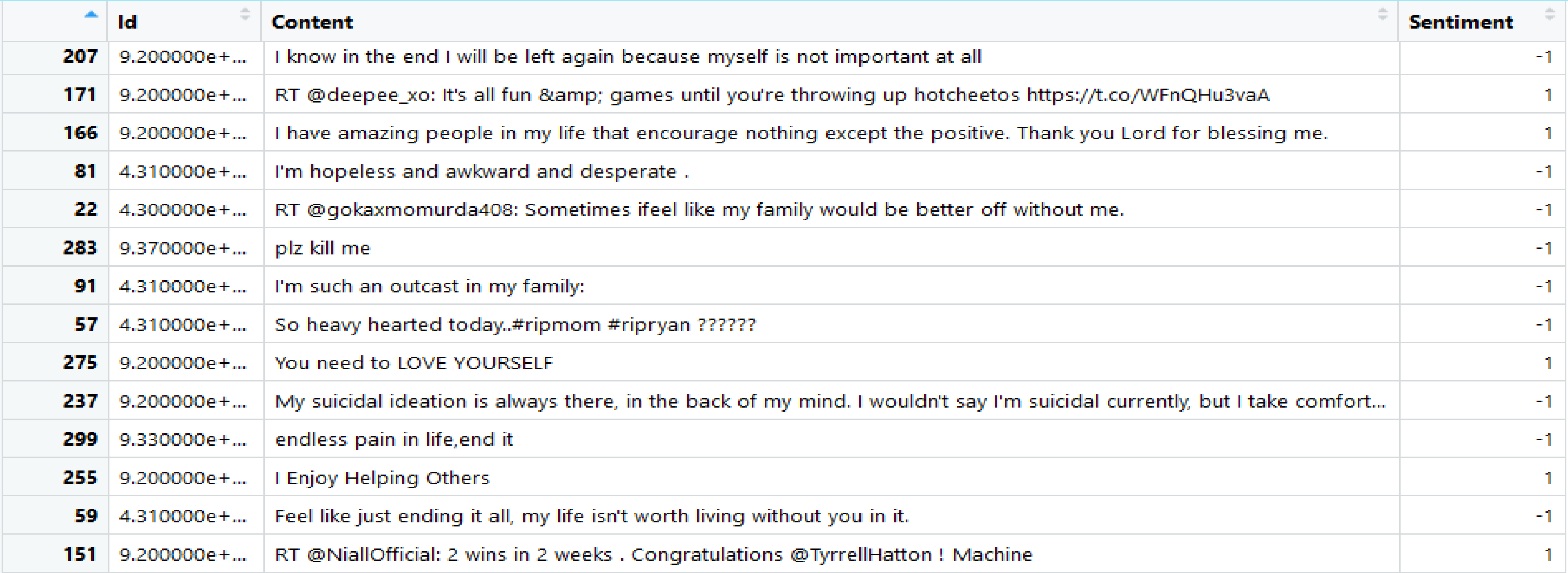
- To detect suicidal ideation by processing the post uploaded by users on the Internet.
- Comparison of supervised learning algorithms for text classification



Proposed Model

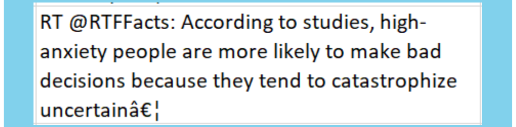


Data Collection



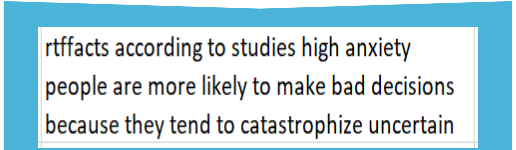
	Id	Content	Sentiment
207	9.200000e+...	I know in the end I will be left again because myself is not important at all	-1
171	9.200000e+...	RT @deepee_xo: It's all fun &amp; games until you're throwing up hotcheetos https://t.co/WFnQHu3vaA	1
166	9.200000e+...	I have amazing people in my life that encourage nothing except the positive. Thank you Lord for blessing me.	1
81	4.310000e+...	I'm hopeless and awkward and desperate .	-1
22	4.300000e+...	RT @gokaxmomurda408: Sometimes ifeel like my family would be better off without me.	-1
283	9.370000e+...	plz kill me	-1
91	4.310000e+...	I'm such an outcast in my family:	-1
57	4.310000e+...	So heavy hearted today..#ripmom #ripryan ??????	-1
275	9.200000e+...	You need to LOVE YOURSELF	1
237	9.200000e+...	My suicidal ideation is always there, in the back of my mind. I wouldn't say I'm suicidal currently, but I take comfort...	-1
299	9.330000e+...	endless pain in life,end it	-1
255	9.200000e+...	I Enjoy Helping Others	1
59	4.310000e+...	Feel like just ending it all, my life isn't worth living without you in it.	-1
151	9.200000e+...	RT @NiallOfficial: 2 wins in 2 weeks . Congratulations @TyrrellHatton ! Machine	1

Data Preprocessing & Feature Extraction



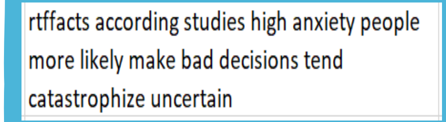
RT @RTFFacts: According to studies, high-anxiety people are more likely to make bad decisions because they tend to catastrophize uncertainâ€¦

Original Tweets



rtffacts according to studies high anxiety people are more likely to make bad decisions because they tend to catastrophize uncertain

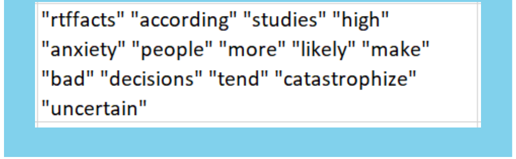
Data Cleaning



rtffacts according studies high anxiety people more likely make bad decisions tend catastrophize uncertain

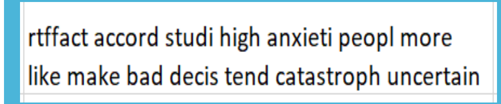
Removal of Stopwords

Tokenization



"rtffacts" "according" "studies" "high" "anxiety" "people" "more" "likely" "make" "bad" "decisions" "tend" "catastrophize" "uncertain"

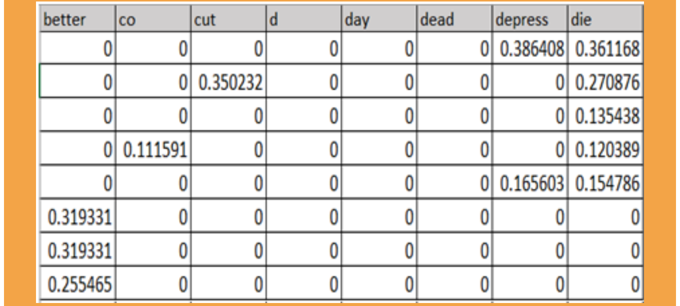
Stemming



rtffact accord studi high anxieti peopl more like make bad decis tend catastroph uncertain

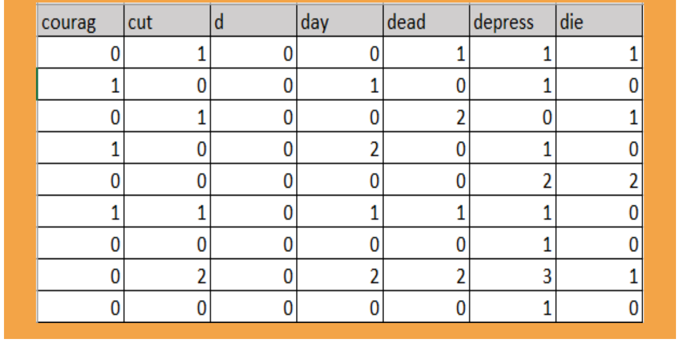
Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF)



better	co	cut	d	day	dead	depress	die
0	0	0	0	0	0	0.386408	0.361168
0	0	0.350232	0	0	0	0	0.270876
0	0	0	0	0	0	0	0.135438
0	0.111591	0	0	0	0	0	0.120389
0	0	0	0	0	0	0.165603	0.154786
0.319331	0	0	0	0	0	0	0
0.319331	0	0	0	0	0	0	0
0.255465	0	0	0	0	0	0	0

Term Document Matrix (TDM)



courag	cut	d	day	dead	depress	die
0	1	0	0	1	1	1
1	0	0	1	0	1	0
0	1	0	0	2	0	1
1	0	0	2	0	1	0
0	0	0	0	0	2	2
1	1	0	1	1	1	0
0	0	0	0	0	1	0
0	2	0	2	2	3	1
0	0	0	0	0	1	0

Algorithms Used

Naïve Bayes

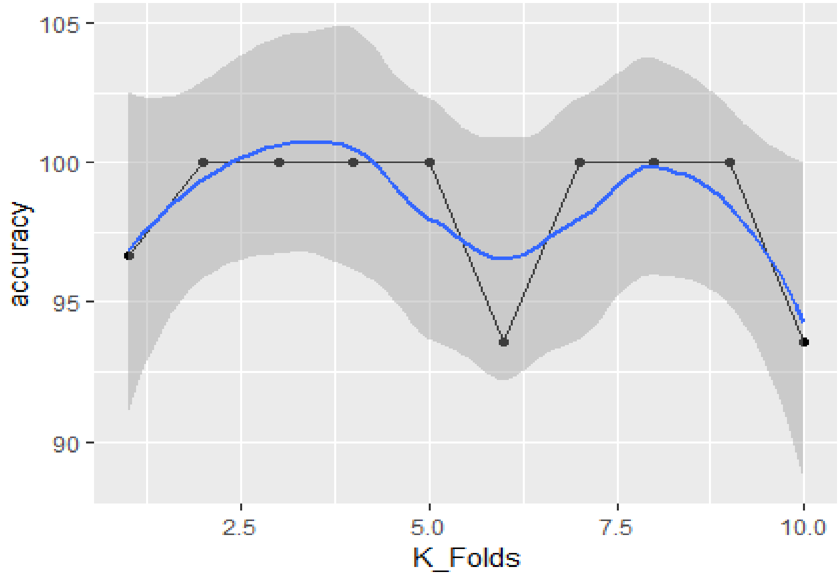
(Maximum a posteriori)

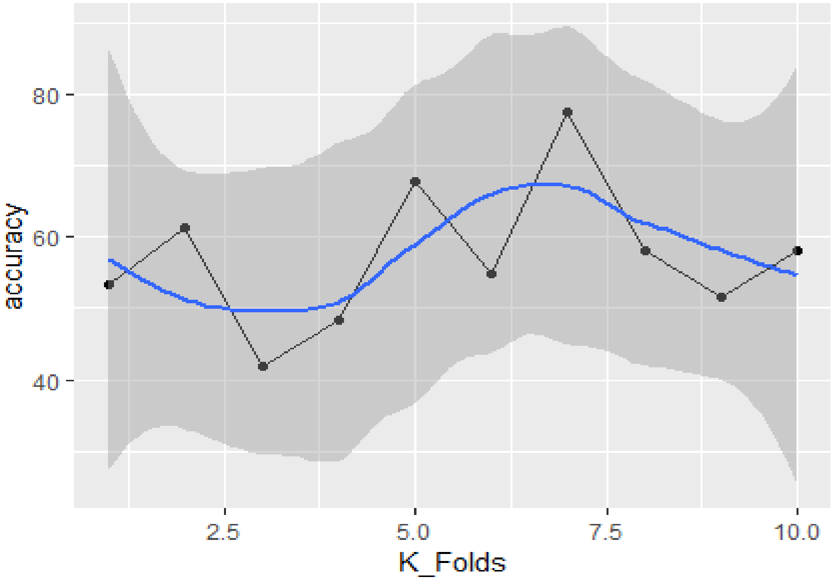
- Simple, fast and easy to implement.
- Highly scalable and leads to good performance.
- Takes features as bag of words.

Confusion Matrix			
		Suicide	Non-Suicide
TF-IDF	Suicide	18	2
	Non-Suicide	0	11
TDM	Suicide	18	13
	Non-Suicide	1	7

Accuracy Analysis

TF-IDF: 10-fold Accuracy





TDM: 10-fold Accuracy

Algorithms Used

Support Vector Machine

(using Stochastic Gradient Descent)

- » Well suited for high dimensional and large amount of data
- » Fast computation
- » High accuracy

Confusion Matrix			
		Suicide	Non-Suicide
TF-IDF	Suicide	20	1
	Non-Suicide	3	7
TDM	Suicide	16	7
	Non-Suicide	1	7

Algorithms Used

Nearest Neighbor

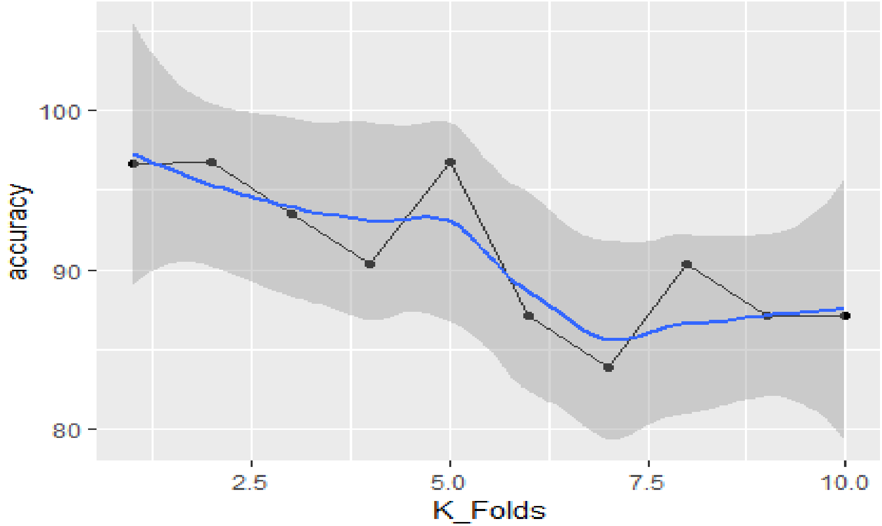
(Rocchio Algorithm)

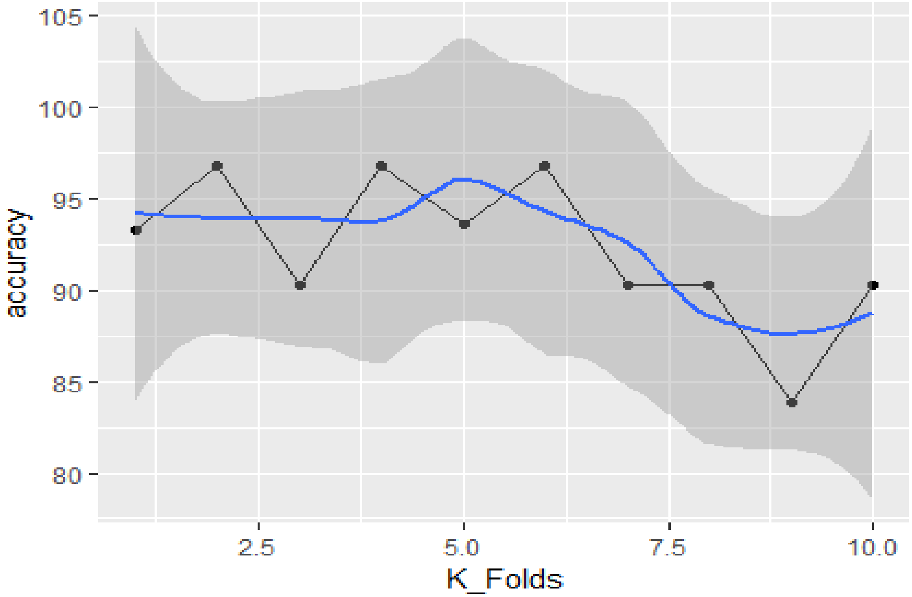
- Useful for non-linear data.
- Relatively high accuracy
- Easy to understand and interpret

Confusion Matrix			
TF-IDF		Suicide	Non-Suicide
	Suicide	174	19
	Non-Suicide	9	107
TDM	Suicide	137	23
	Non-Suicide	40	109

Accuracy Analysis

TF-IDF: 10-fold Accuracy



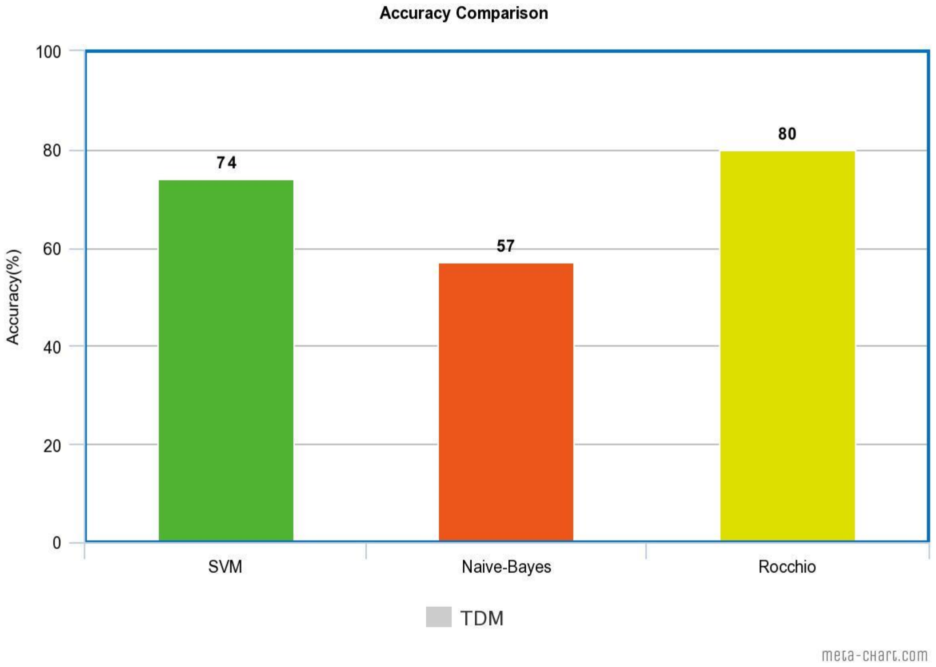


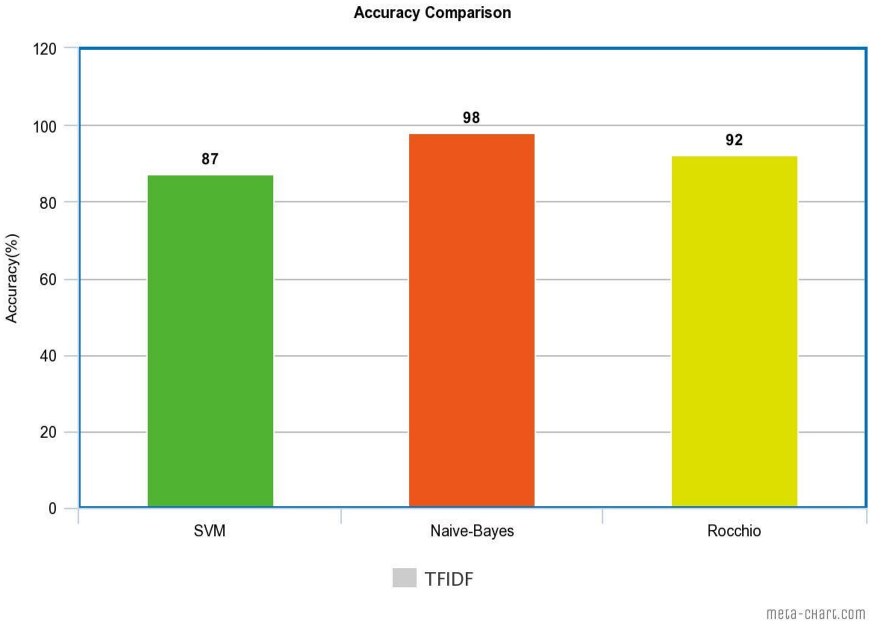
TDM: 10-fold Accuracy

Performance Comparison

Classifiers	Features	Accuracy	Precision	Recall	Specificity	F1 Score
Naïve Bayes	TF-IDF	98.3	90.00	100.00	84.61538	94.736
	TDM	57.2	58.064	100.00	0	73.469
Support Vector Machine (SVM)	TF-IDF	87.096	95.238	86.956	86.956	90.909
	TDM	74.193	69.565	94.117	50.000	80.000
Rocchio Algorithm	TF-IDF	91.0	90.155	95.081	84.920	92.553
	TDM	79.6	85.625	77.401	82.575	81.305

Performance Comparison





Conclusion

- Naïve-Bayes gave the best performance for TF-IDF feature set. It reaffirms the importance of probabilistic view for text classification.
- Accuracies for TF-IDF outperforms that for TDM feature set indicating that relative term frequency and Inverse document frequency is a better measure of similarities in texts.
- As per the results, count of False-Negatives is less than False-Positives for most of the algorithms indicating lesser count of suicidal tweets classified as non-suicidal.

Research Papers Referred

- Supervised Learning for Suicidal Ideation Detection in Online User Content Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long, “Supervised Learning for Suicidal Ideation Detection in Online User Content,” Complexity, vol. 2018, Article ID 6157249, 10 pages, 2018. https://doi.org/10.1155/2018/6157249
- P. Burnap, W. Colombo, and J. Scourfield. Machine classification and analysis of suicide-related communication on Twitter. In Proceedings of the 26th ACM Conference on Hypertext & Social Media, pages 75–84. ACM, 2015
- Birjali, M., Beni-hssane, A., & MohammedErritali (2016). Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms.

THANK YOU!

Any questions???