CSE 454 Data Mining Final Project Report

Ahmed Semih Özmekik

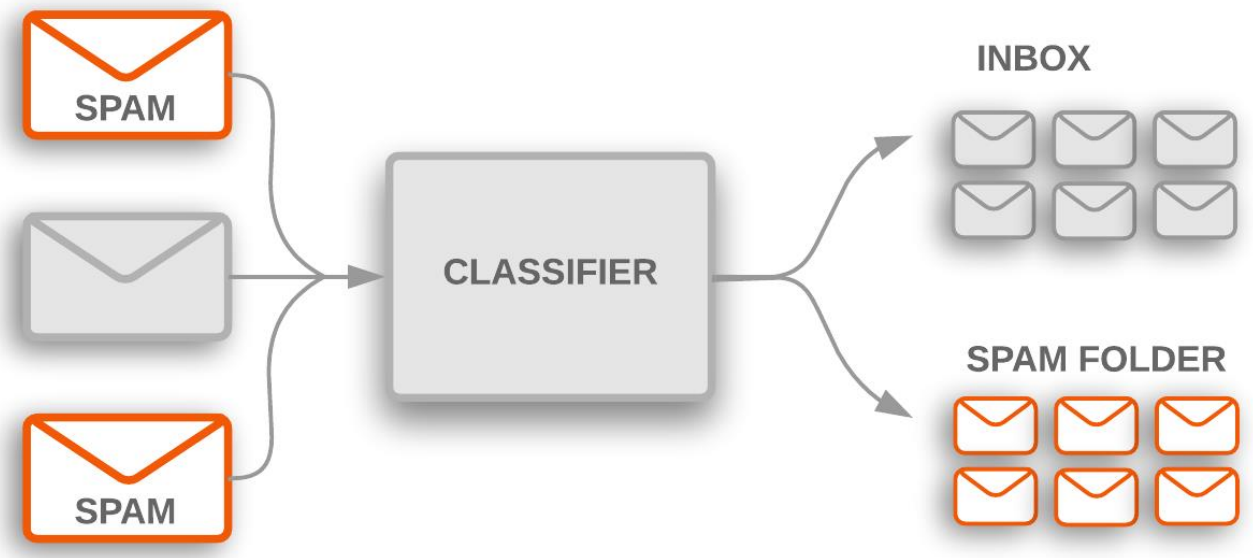
Project Definition

- Our main goal in the project was to use many data mining algorithms in the literature that can be used at the point where we will specialize by applying many pre-processing, post-processing methods related to data mining.
- Based on this, we first determined a paper.[1] Then, we focused on improving the results obtained there by examining the experimental study in the paper. This project was firstly an implementation of a paper, and then by doing different studies on the dataset suggested in this paper, it was aimed and achieved better scores than the scores obtained in this paper.

Project Definition

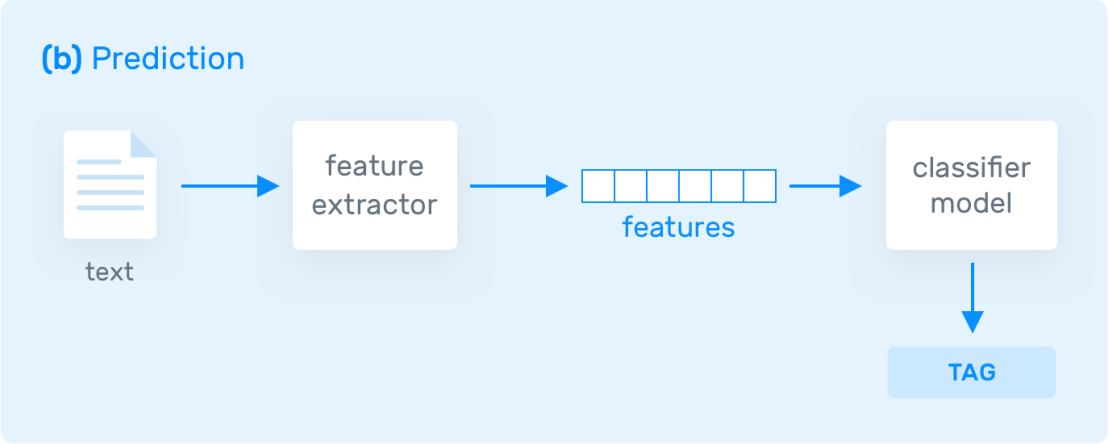
- The subject is text classification. There are many different and successful methods in the current literature on this subject. We present the results of 8 different methods. For each method, there are 6 different preprocessing parameters, 2 different feature vector method parameters, and 5 different feature number definition parameters for the feature selection method. In other words, 8x2x5 = 80 different results were obtained for each method, hence 80x8 = 640 different results obtained in total.
- For these 480 different results, by making some comparisons between them, the results of all methods and independent variables in the experiment were analyzed, and the dependent variable f1 score was observed. As a result, all data were recorded and the most successful parameter selection was determined.

Text Classification



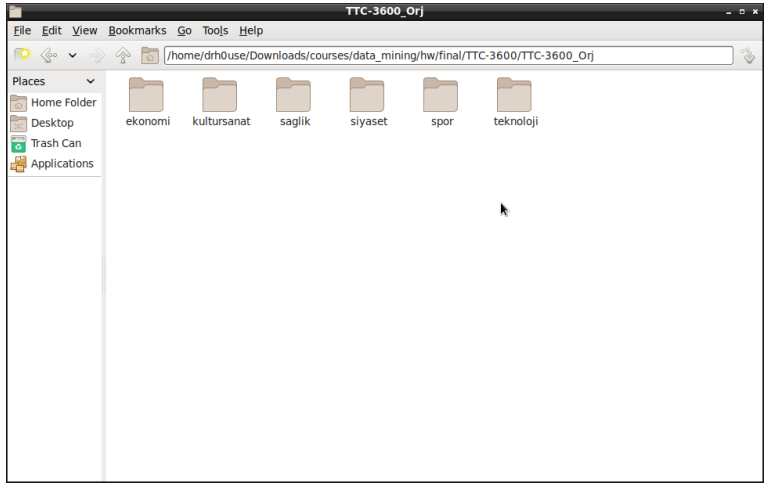
Text Classification

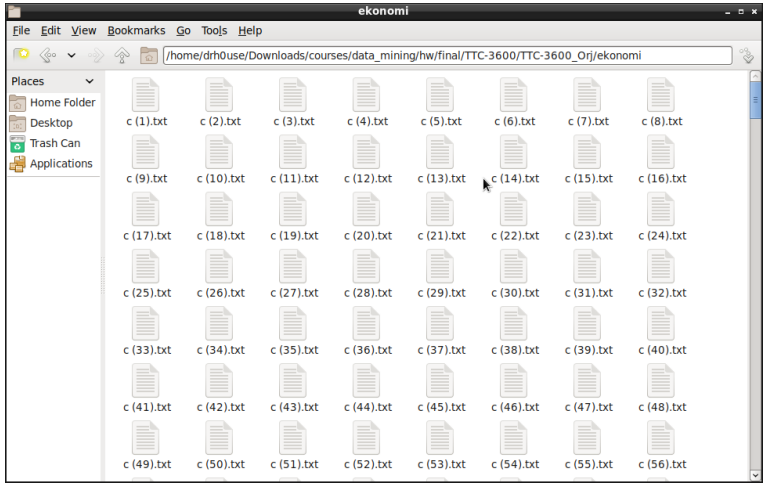
- Importance of pre-processing.
- Post-processing as a feature extraction.



Dataset: TTC-3600

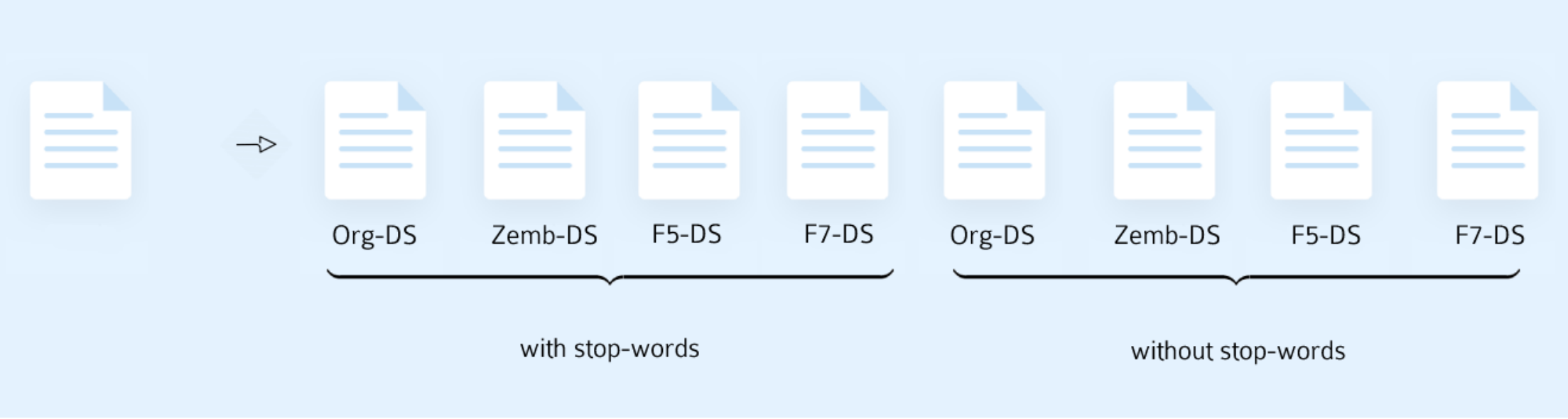
- The dataset includes a total of 3600 documents, 600 of them from each class (economy, culture-arts, health, politics, sports and technology), all collected from well-known and known news portals.
- It was taken from the RSS feeds of "Hurriyet, Posta, Iha, HaberTurk, Radikal and Zaman" news sites between May-July 2015 and parsed.





Pre-processing

- Preprocessing is the most important process of TC subject. At this stage, we have applied several methods on our data set, usually based on removal, that is, to remove noise on the data.
- Stages: Number removal, lower-casing, punctutation removal, normalizing etc.

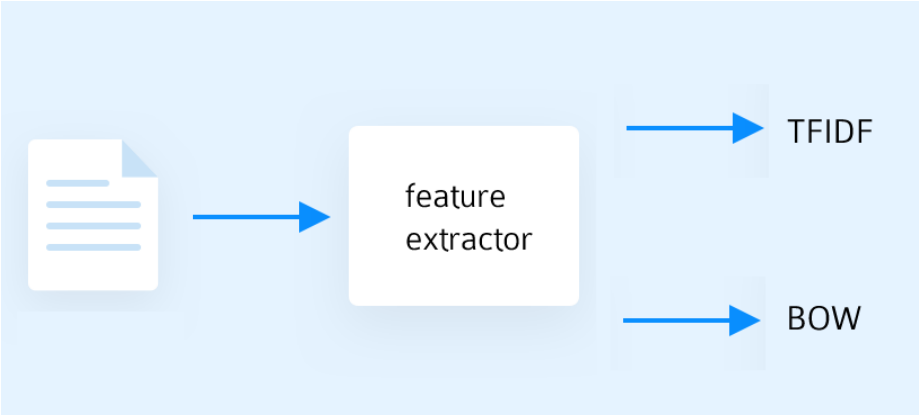


Pre-processing

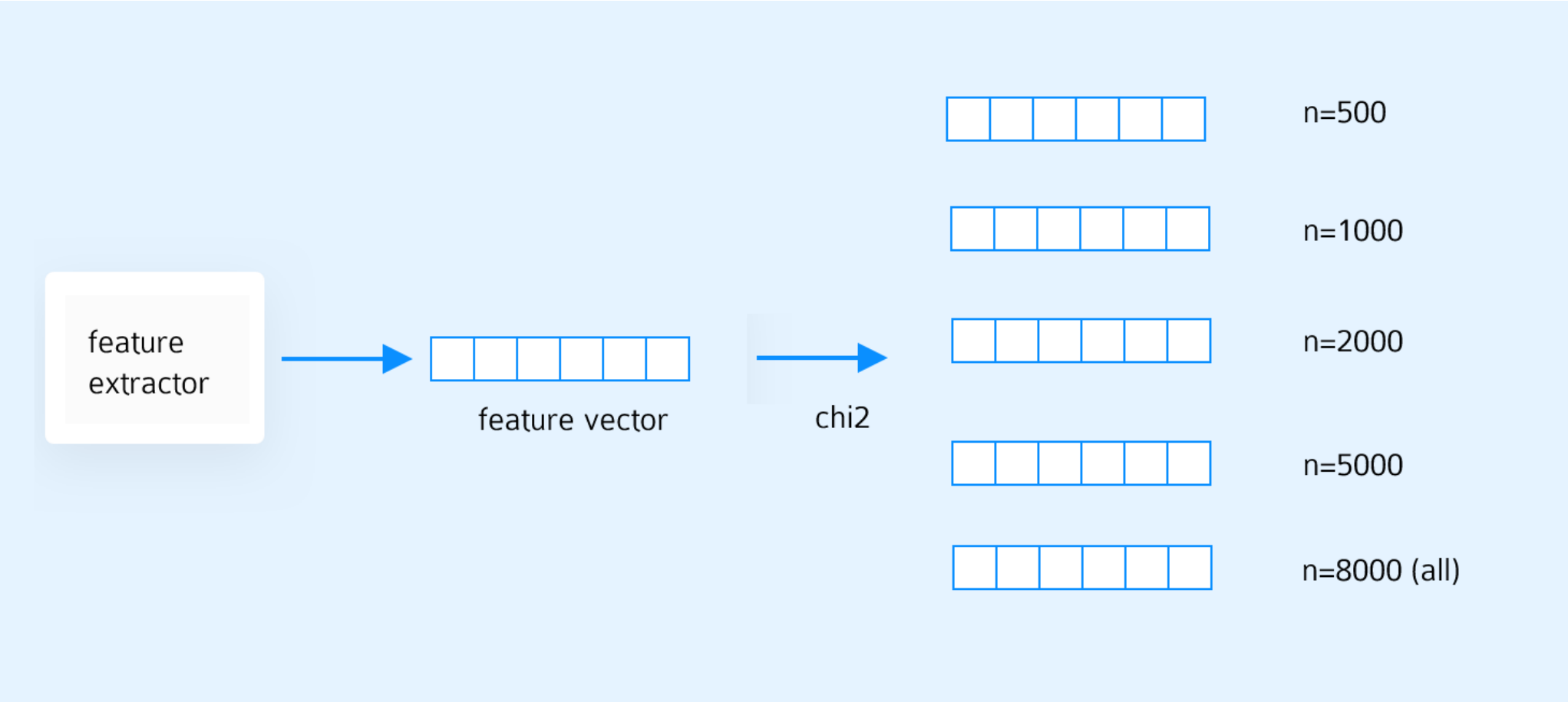
Dataset name	Stop word filtering	Stemmer
Original-DS	No	No stemmer
F5-DS	No	FPS-5
F7-DS	No	FPS-7
Zemb-DS	No	Zemberek
OriginalSW-DS	Yes	No Stemmer
F5SW-DS	Yes	FPS-5
F7SW-DS	Yes	FPS-7
ZembSW-DS	Yes	Zemberek

Post-processing: Feature Extraction

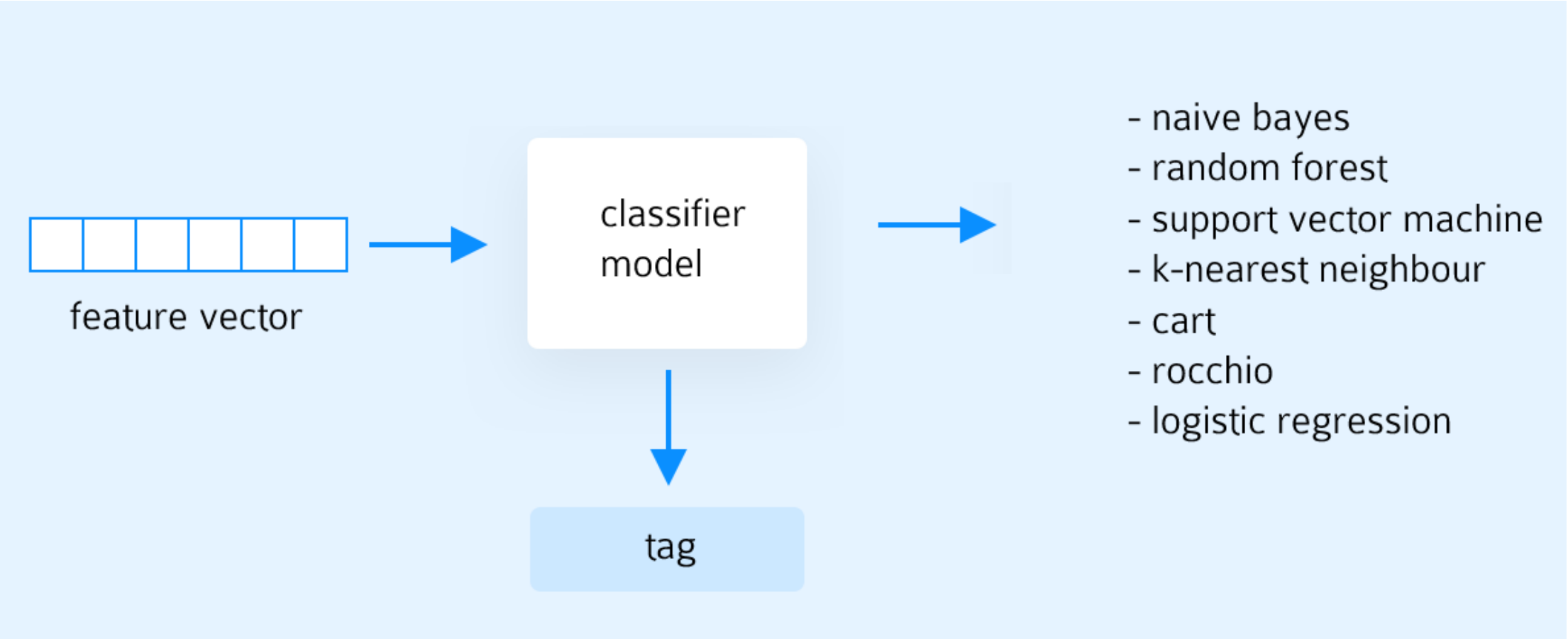
- The first category represents the different methods that can be applied in extracting a feature vector from the texts we have. We chose two approaches in the literature to use in our experiments: BOW (Bag of words) and TFIDF (term frequency–inverse document frequency) vectorization methods.



Post-processing: Feature Selection



Methods


feature vector
classifier model
tag
- naive bayes
- random forest
- support vector machine
- k-nearest neighbour
- cart
- rocchio
- logistic regression

Experiments: Naïve Bayes

			Dataset	Score
WITH STOPWORDS	TFIDF	500	OrgDS	0.8903
			ZembDS	0.9139
			F5DS	0.9144
			F7DS	0.9158
		1000	OrgDS	0.9089
			ZembDS	0.9264
			F5DS	0.9225
			F7DS	0.9233
		2000	OrgDS	0.9208
			ZembDS	0.9300
			F5DS	0.9303
			F7DS	0.9297
		5000	OrgDS	0.9261
			ZembDS	0.9339
			F5DS	0.9319
			F7DS	0.9322
		8000 (all)	OrgDS	0.9256
			ZembDS	0.9342
			F5DS	0.9306
			F7DS	0.9328
	BOW	500	OrgDS	0.8914
			ZembDS	0.9164
			F5DS	0.9136
			F7DS	0.9150
		1000	OrgDS	0.9083
			ZembDS	0.9267
			F5DS	0.9211
			F7DS	0.9194
		2000	OrgDS	0.9178
			ZembDS	0.9294
			F5DS	0.9267
			F7DS	0.9283
		5000	OrgDS	0.9289
			ZembDS	0.9386
			F5DS	0.9322
			F7DS	0.9292
		8000 (all)	OrgDS	0.9297
			ZembDS	0.9378
			F5DS	0.9322
			F7DS	0.9314

Experiments: Random Forest

			Dataset	Score
WITH STOPWORDS	TFIDF	500	OrgDS	0.8783
			ZembDS	0.9033
			F5DS	0.9083
			F7DS	0.8950
		1000	OrgDS	0.8825
			ZembDS	0.9089
			F5DS	0.9114
			F7DS	0.9000
		2000	OrgDS	0.8875
			ZembDS	0.9108
			F5DS	0.9114
			F7DS	0.9017
		5000	OrgDS	0.8958
			ZembDS	0.9150
			F5DS	0.9164
			F7DS	0.9081
		8000 (all)	OrgDS	0.8903
			ZembDS	0.9133
			F5DS	0.9122
			F7DS	0.9108
	BOW	500	OrgDS	0.8658
			ZembDS	0.9025
			F5DS	0.9042
			F7DS	0.8944
		1000	OrgDS	0.8769
			ZembDS	0.9075
			F5DS	0.9047
			F7DS	0.8994
		2000	OrgDS	0.8856
			ZembDS	0.9083
			F5DS	0.9100
			F7DS	0.9053
		5000	OrgDS	0.8944
			ZembDS	0.9114
			F5DS	0.9131
			F7DS	0.9086
		8000 (all)	OrgDS	0.8950
			ZembDS	0.9122
			F5DS	0.9136
			F7DS	0.9078

Experiments: Support Vector Machine

			Dataset	Score
WITH STOPWORDS	TFIDF	500	OrgDS	0.8969
			ZembDS	0.9236
			F5DS	0.9222
			F7DS	0.9211
		1000	OrgDS	0.9211
			ZembDS	0.9361
			F5DS	0.9361
			F7DS	0.9356
		2000	OrgDS	0.9300
			ZembDS	0.9442
			F5DS	0.9433
			F7DS	0.9386
		5000	OrgDS	0.9356
			ZembDS	0.9492
			F5DS	0.9475
			F7DS	0.9458
		8000 (all)	OrgDS	0.9389
			ZembDS	0.9508
			F5DS	0.9506
			F7DS	0.9483
	BOW	500	OrgDS	0.8731
			ZembDS	0.8844
			F5DS	0.8869
			F7DS	0.8833
		1000	OrgDS	0.8772
			ZembDS	0.9011
			F5DS	0.8975
			F7DS	0.8925
		2000	OrgDS	0.8806
			ZembDS	0.9094
			F5DS	0.9081
			F7DS	0.8994
		5000	OrgDS	0.8942
			ZembDS	0.9192
			F5DS	0.9122
			F7DS	0.9025
		8000 (all)	OrgDS	0.9050
			ZembDS	0.9208
			F5DS	0.9186
			F7DS	0.9156

Experiments: K-Nearest Neighbour

			Dataset	Score
WITH STOPWORDS	TFIDF	500	OrgDS	0.7861
			ZembDS	0.8367
			F5DS	0.8261
			F7DS	0.8189
		1000	OrgDS	0.7394
			ZembDS	0.7744
			F5DS	0.7864
			F7DS	0.7567
		2000	OrgDS	0.6092
			ZembDS	0.6886
			F5DS	0.6783
			F7DS	0.6442
		5000	OrgDS	0.7356
			ZembDS	0.8569
			F5DS	0.8244
			F7DS	0.7997
		8000 (all)	OrgDS	0.8767
			ZembDS	0.9003
			F5DS	0.9017
			F7DS	0.9006
	BOW	500	OrgDS	0.6311
			ZembDS	0.7447
			F5DS	0.7594
			F7DS	0.7100
		1000	OrgDS	0.5994
			ZembDS	0.7242
			F5DS	0.7319
			F7DS	0.6808
		2000	OrgDS	0.5744
			ZembDS	0.6914
			F5DS	0.6978
			F7DS	0.6431
		5000	OrgDS	0.5200
			ZembDS	0.6392
			F5DS	0.6244
			F7DS	0.5581
		8000 (all)	OrgDS	0.4853
			ZembDS	0.6222
			F5DS	0.5858
			F7DS	0.5386

Experiments: CART

			Dataset	Score
WITH STOPWORDS	TFIDF	500	OrgDS	0.7850
			ZembDS	0.8072
			F5DS	0.8061
			F7DS	0.7944
		1000	OrgDS	0.7592
			ZembDS	0.7897
			F5DS	0.8044
			F7DS	0.7944
		2000	OrgDS	0.7472
			ZembDS	0.7894
			F5DS	0.7936
			F7DS	0.7869
		5000	OrgDS	0.7244
			ZembDS	0.7806
			F5DS	0.7856
			F7DS	0.7747
		8000 (all)	OrgDS	0.7339
			ZembDS	0.7842
			F5DS	0.7867
			F7DS	0.7753
	BOW	500	OrgDS	0.7503
			ZembDS	0.7858
			F5DS	0.7769
			F7DS	0.7925
		1000	OrgDS	0.7503
			ZembDS	0.7808
			F5DS	0.7786
			F7DS	0.7853
		2000	OrgDS	0.7592
			ZembDS	0.7794
			F5DS	0.7753
			F7DS	0.7850
		5000	OrgDS	0.7544
			ZembDS	0.7803
			F5DS	0.7761
			F7DS	0.7769
		8000 (all)	OrgDS	0.7500
			ZembDS	0.7831
			F5DS	0.7817
			F7DS	0.7811

Experiments: Rocchio

			Dataset	Score
WITH STOPWORDS	TFIDF	500	OrgDS	0.8342
			ZembDS	0.8872
			F5DS	0.8761
			F7DS	0.8700
		1000	OrgDS	0.8608
			ZembDS	0.9019
			F5DS	0.8944
			F7DS	0.8892
		2000	OrgDS	0.8781
			ZembDS	0.9053
			F5DS	0.9025
			F7DS	0.8981
		5000	OrgDS	0.8931
			ZembDS	0.9106
			F5DS	0.9081
			F7DS	0.9092
		8000 (all)	OrgDS	0.8958
			ZembDS	0.9106
			F5DS	0.9106
			F7DS	0.9111
	BOW	500	OrgDS	0.4389
			ZembDS	0.5375
			F5DS	0.6003
			F7DS	0.5264
		1000	OrgDS	0.4467
			ZembDS	0.5406
			F5DS	0.6039
			F7DS	0.5275
		2000	OrgDS	0.4578
			ZembDS	0.5450
			F5DS	0.6053
			F7DS	0.5358
		5000	OrgDS	0.4661
			ZembDS	0.5472
			F5DS	0.6083
			F7DS	0.5439
		8000 (all)	OrgDS	0.4667
			ZembDS	0.5489
			F5DS	0.6086
			F7DS	0.5450

Experiments: Logistic Regression

			Dataset	Score
WITH STOPWORDS	TFIDF	500	OrgDS	0.8861
			ZembDS	0.9158
			F5DS	0.9142
			F7DS	0.9086
		1000	OrgDS	0.9081
			ZembDS	0.9297
			F5DS	0.9253
			F7DS	0.9231
		2000	OrgDS	0.9219
			ZembDS	0.9367
			F5DS	0.9342
			F7DS	0.9328
		5000	OrgDS	0.9278
			ZembDS	0.9428
			F5DS	0.9392
			F7DS	0.9367
		8000 (all)	OrgDS	0.9300
			ZembDS	0.9433
			F5DS	0.9417
			F7DS	0.9389
	BOW	500	OrgDS	0.8839
			ZembDS	0.9011
			F5DS	0.8994
			F7DS	0.8964
		1000	OrgDS	0.8950
			ZembDS	0.9147
			F5DS	0.9119
			F7DS	0.9097
		2000	OrgDS	0.9003
			ZembDS	0.9189
			F5DS	0.9203
			F7DS	0.9178
		5000	OrgDS	0.9094
			ZembDS	0.9247
			F5DS	0.9244
			F7DS	0.9225
		8000 (all)	OrgDS	0.9111
			ZembDS	0.9275
			F5DS	0.9269
			F7DS	0.9247

Conclusion

- Based on the results of the above deep and extensive experiments, we tried to determine what is the best parameter for a method, and then to determine the common point between these parameters.
- Common Patterns;
- How did removal of stop-words effect the results?
- How did feature extraction method effect the results?
- Among the best 7 scores, 4 are the result obtained on Zemb-DS. Two of them were obtained on F5-DS and the last one on F7-DS.

Conclusion

- The best results, together with these common patterns we have caught with; using the TFIDF in the attribute vector, the preeprocessed data set is stuck in the resulting range, with the stop-words not extracted. So now, in order to gain a more distant view of the picture, we can draw the following table, which only takes this range into our perspective.

Conclusion

	Feature: 500				Feature: 5000				Feature: 8000 (All)			
	OrgDS	ZembDS	F5DS	F7DS	OrgDS	ZembDS	F5DS	F7DS	OrgDS	ZembDS	F5DS	F7DS
NB	0.8903	0.9139	0.9144	0.9158	0.9261	0.9339	0.9319	0.9322	0.9256	0.9342	0.9306	0.9328
RF	0.8783	0.9033	0.9083	0.8950	0.8958	0.9150	0.9164	0.9081	0.8903	0.9133	0.9122	0.9108
SVM	0.8969	**0.9236**	0.9222	0.9211	0.9356	**0.9492**	0.9475	0.9458	0.9389	0.9508	**0.9506**	0.9483
KNN	0.7861	0.8367	0.8261	0.8189	0.7356	0.8569	0.8244	0.7997	0.8767	0.9003	0.9017	0.9006
CART	0.7850	0.8072	0.8061	0.7944	0.7244	0.7806	0.7856	0.7747	0.7339	0.7842	0.7867	0.7753
Rocchio	0.8342	0.8872	0.8761	0.8700	0.8931	0.9106	0.9081	0.9092	0.8958	0.9106	0.9106	0.9111
LR	0.8861	0.9158	0.9142	0.9086	0.9278	0.9428	0.9392	0.9367	0.9300	0.9433	0.9417	0.9389

Conclusion

	[1]				Ours			
	OrgDS	ZembDS	F5DS	F7DS	OrgDS	ZembDS	F5DS	F7DS
NB	0.8294	0.8719	0.8222	0.8403	0.9256	**0.9342**	0.9319	0.9328
RF	0.8887	0.9103	0.8828	0.8859	0.8903	**0.9150**	0.9122	0.9108
SVM	0.8603	0.8497	0.8239	0.8356	0.9389	0.9492	**0.9506**	0.9483
KNN	0.7311	0.7497	0.6944	0.7256	0.8767	0.9003	**0.9017**	0.9006
CART	0.7897	0.7939	0.7736	0.7597	0.7850	0.8072	**0.8061**	0.7944

References

- [1] Kilinç, Deniz & Ozcift, Akin & Bozyiğit, Fatma & Yildirim, Pelin & Yucalar, Fatih & Borandağ, Emin. (2017). TTC-3600: A new benchmark dataset for Turkish text categorization. Journal of Information Science. 43. 174-185.