PREDICTION MODELS BASED ON MAX-STEMS

(or harnessing imbalanced data)

Chapter Two: A Combinatorial Approach

Ahmet Furkan EMREHAN

(matahmet@gmail.com)

MOTIVATION

- This study is an extension of previous study (chapter one) with combinatiorial approach.
- In chapter one, I examine five models using distribution of stems separately. Combination of stems with s-elements have a potential to help efficient prediction. Because documents including comination of stems may be semantically closer than one stem based prediction (defined in chapter one).

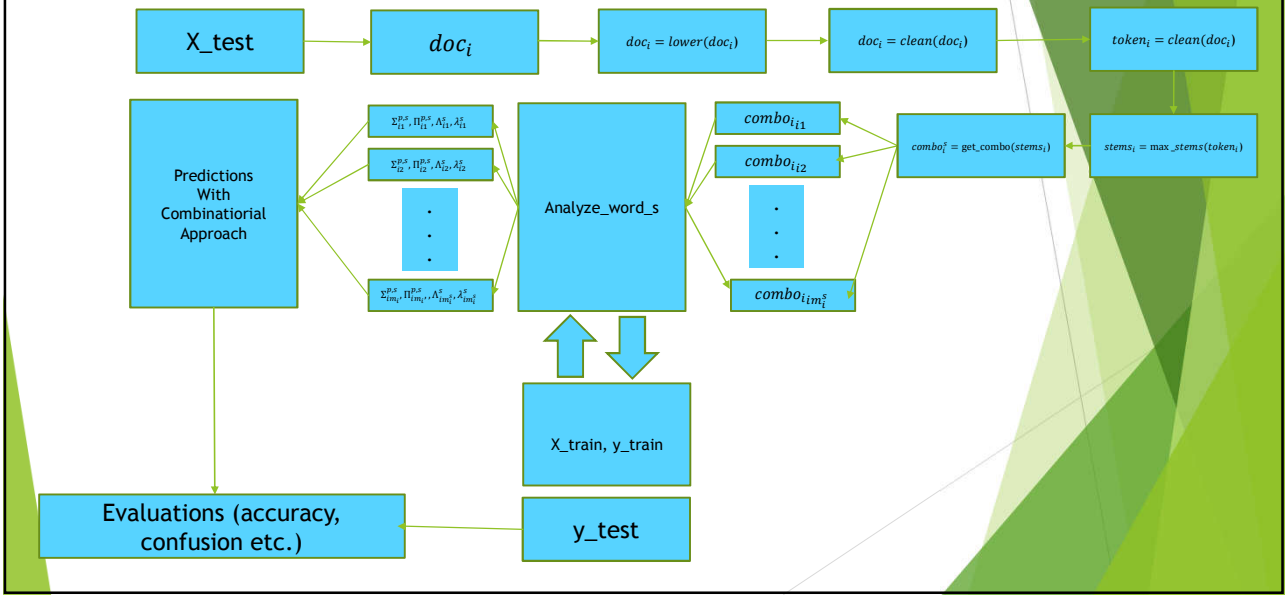
ADAPTATION OF COMPONENTS OF MODELS

- $p, Label^p, n, doc_i \text{ and } \Sigma^p$, components of general parameters, are defined in Chapter One *Slide 8*
- $combo_{ij}^s$: *combination of stems indexed j wit s elements of* doc_i
  (*stem can be c osen as max stem mentioned previous slides.*)
- m_i^s: *counts of* $combo_{ij}^s$
- $\Sigma_{ij}^{p,s}$: *counts of documents, w ic include all elements of* $combo_{ij}^s$, *labelled wit category wit index p in train set*
- $\Lambda_{ij}^s := Label^q$ *w ere* $q = \arg\max_p \Sigma_{ij}^{p,s}$
- $\Lambda_i^{p,s}$: counts of Λ_{ij}^s *w ic equals to* $Label^p$

ADAPTATION OF COMPONENTS OF MODELS

- λ_{ij}^s: *sum of lengt of stems in* $combo_{ij}^s$
- $\rho_i^{p,s} \quad \frac{\sum_{j=1}^{m_i} \Sigma_{ij}^{p,s}}{\Sigma^p}$ *
- *in case t at* $\Sigma^p = 0, \rho_i^p = 0$
- $\Pi_{ij}^{p,s} \quad \frac{\Sigma_{ij}^{p,s}}{\sum_{q=1}^{n} \Sigma_{ij}^{q}}$
  (*it can be considered as probability of* $combo_{ij}^s$, *labelled wit category wit p index*)
- $\overline{\Pi_i^{p,s}} := average_j \left(\Pi_{ij}^{p,s}\right)$*suc t at all* "j "*s meet t e condition* $\Pi_{ij}^{p,s} > 0$ *
  in case that $\Sigma_{ij}^{p,s} = 0$ *for all* $p = 1,2, \dots n, \overline{\Pi_i^{p,s}} = 0$
- $\widehat{\Pi_i^{p,s}} \quad \max_j (\Pi_{ij}^{p,s})$

General Scheme for Prediction Models with Combinatorial Approach



Model 1

- $predict\_cb_1(doc_i, s) = \begin{cases} Label^q & if\ \Sigma_{ij}^{p,s} = 0 < \Sigma_{ij}^{q,s}\ for\ all\ p \neq q \\ Label^q\ w\ ere\ q = \arg 2nd\ max_p\ \Lambda_i^{p,s} ot\ \ erwise \end{cases}$

in case t at q is not uniqe, q is c osen as t e minimum index meeting t e condition

Model 2

- $predict\_cb_2(doc_i, s) = Label^q \; w \;\; ere \; q = \arg max_p \, \rho_i^{p,s}$

Model 3

- $predict\_cb_3(doc_i, s) = Label^q \; w \;\; ere \; q = \arg max_p \, \overline{\Pi_i^{p,s}}$

Model 4

- $predict\_cb_4(doc_i, s) = Label^q$ *w ere* $q = \arg max_p \widehat{\Pi_i^{p,s}}$
  in case t at q is not uniqe, q is c osen as t e minimum index meeting t e condition

Model 5

- $predict\_cb_5(doc_i, s) = Label^q$ *w ere* $q = \arg max_p \left(max_j \left(\lambda_{ij} \ \frac{\Sigma_{ij}^p}{\Sigma^p}\right)\right)$

A Trivial Result

- $predict\_cb_k(doc_i, 1) = predict_k(doc_i)\ for\ k = 1,2,\dots,5$
- Moreover all parameters in case $s = 1$, equal to corresponding parameters in chapter one (slides 8-10) For example $m_i^s = m_i, \Sigma_{ij}^{p,1} = \Sigma_{ij}^p$ and $\overline{\Pi_i^{p,1}} = \overline{\Pi_i^p}$.

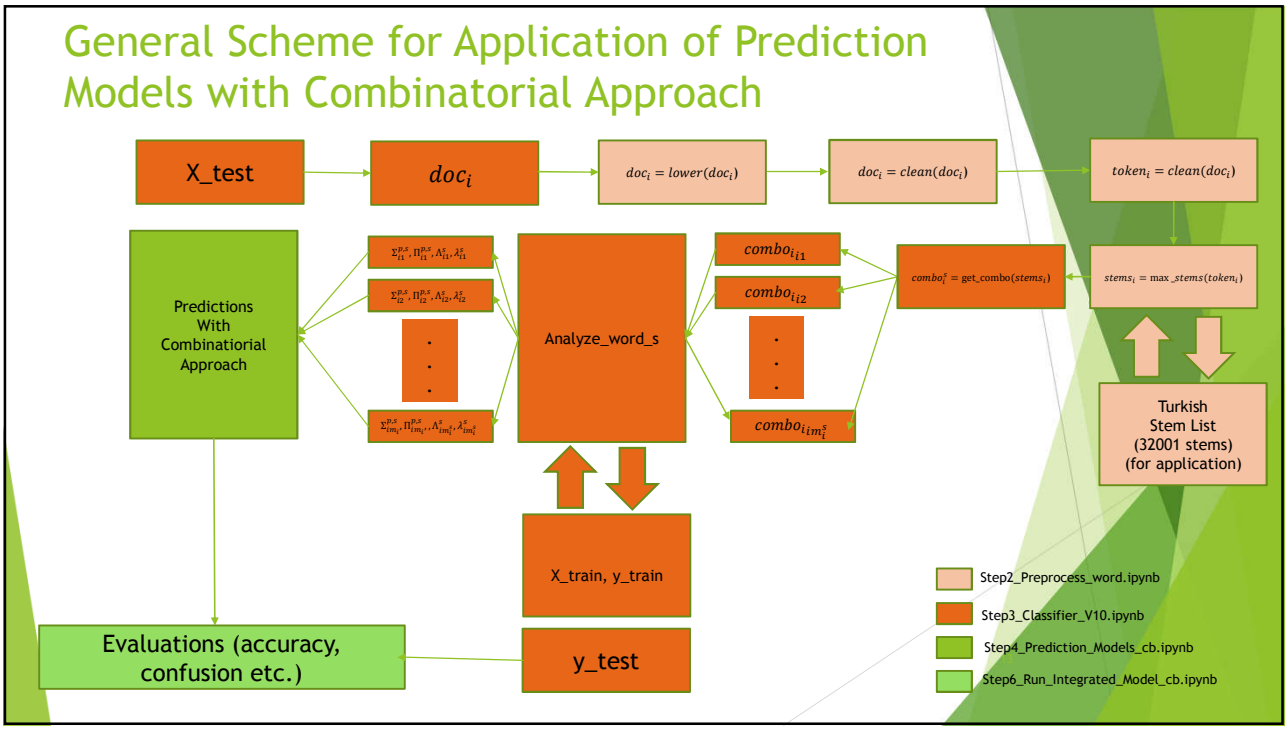
Application (introduction)

- We use data of «nayn.co» a news portal in Turkish Language. Data is imported by url «"https://raw.githubusercontent.com/naynco/nayn.data/master/classification_clean.csv"» as done in chapter one.
- Head of data is presented below

	Title	Categories
12006	58 Saniyede Katar Meselesi? Katar krizi nedir?...	DÜNYA
12496	58 Saniyede Türkiye - Almanya Gerginliği	DÜNYA
12877	Adriana Lima, Bomba Aşkla İlgili İlk Kez Konuş...	DÜNYA
12878	Galatasaraylı Taraftarlar Patladı: İstifa Edin	SPOR
12880	Galatasaray'dan Ayrılan Sabri, Neredeyse Bedav...	SPOR

There are 11622 documents («Title» column) with label («DÜNYA» (World), «SPOR» (Sports), «SANAT»(Art) and «Teknoloji»(Technology)). But data is imbalanced in favor of category «DÜNYA» such that the counts [and percentages] of categories 9226[%79] ,1967 [%17],285 [%2] and 144 [%1] respectively.

General Scheme for Application of Prediction Models with Combinatorial Approach



Application (computations)

- Pandas and Sklearn libraries in Python is used for application of methods. Test size is chosen as 0.2 and random_state parameter for partition as 57.
- Values of parameters in model are computed below
- Counts of categories: $n = 4$
- Indexes and name of categories: $p = 1,2,3\ and\ 4\ ,\ Label^p$ = "DÜNYA","SPOR","SANAT and "Teknoloji", $respectively$
- Counts of categories in train set : $\Sigma^1 = 7384, \Sigma^2 = 1568, \Sigma^3 = 229\ and\ \Sigma^4 = 116$

Application (computations)

- Now Let's show an example and compute its parameters (or compounds of models). We apply models to document, used in chapter one, with index number $i = 38296$, $rank\ (in\ test\ set) = 1356$ (index number may not be related to rank). We examine case $s = 2$, because set of $combo_{ij}^{s}$ of the document is empty for $s > 2$.
- doc_i: $2\ kedi\ 2\ yıldır\ sanat\ müzesine\ girmeye\ çalışıyor.$ $(en: 2\ cats\ try\ to\ enter\ art\ museum\ for\ 2\ years)$
- $label_i$: $DÜNYA\ (en: world)$
- $stems$: $['kedi', 'yıldır', 'sanat', 'müze', 'gir', 'çalı']$
- Output of analyze_doc(doc,X_train,y_train): $j = 1, \dots, m_i^2 = 7$

```
[[['sanat','müze'],9,[0, 0, 2, 0],[0.0, 0.0, 1.0, 0.0],'SANAT','No Prediction'],
[['kedi','gir'],7,[1, 0, 0, 0],[1.0, 0.0, 0.0, 0.0],'DÜNYA','No Prediction'],
[['müze','çalı'],8,[2, 0, 1, 0],[0.67, 0.0, 0.33, 0.0],'DÜNYA', 'SANAT'],
[['kedi','yıldır'],10,[1, 0, 0, 0],[1.0, 0.0, 0.0, 0.0],'DÜNYA','No Prediction'],
[['gir','çalı'],7,[3, 1, 0, 0],[0.75, 0.25, 0.0, 0.0],'DÜNYA', 'SPOR'],
[['yıldır','çalı'],10,[0, 2, 0, 0],[0.0, 1.0, 0.0, 0.0],'SPOR','No Prediction'],
[['sanat','çalı'],9,[1, 0, 0, 0],[1.0, 0.0, 0.0, 0.0],'DÜNYA','No Prediction']]
```

$combo_{ij}^{2}$ | λ_{ij}^{2} | $\Sigma_{ij}^{p,2}$ | $\Pi_{ij}^{p,2}$ | $\Lambda_{i}^{p,2}$ | Category of 2nd max $\Sigma_{ij}^{p,2}$ (not used)

Application (computations)

- $i = 38296,\ s = 2$
- $combo$: $\text{combo}_{i1}^{2} = ['sanat', 'müze']$, $\text{combo}_{i2}^{2} = ['kedi', 'gir']$, $\text{combo}_{i3}^{2} = ['müze', 'çalı']$, $\text{combo}_{i4}^{2} = ['kedi', 'yıldır']$, $\text{combo}_{i5}^{2} = ['gir', 'çalı']$, $\text{combo}_{i6}^{2} = ['yıldır', 'çalı']$, $\text{combo}_{i7}^{2} = ['sanat', 'çalı']$
- $sum\ of\ lengt\ \ of\ stems\ in\ combo_{ij}^{s}$: $\lambda_{i1}^{2} = 9, \lambda_{i2}^{2} = 7, \lambda_{i3}^{2} = 8, \lambda_{i4}^{2} = 9, \lambda_{i5}^{2} = 7, \lambda_{i6}^{2} = 10, \lambda_{i7}^{2} = 9$
- $\Sigma_{ij}^{p,s}$: $counts\ of\ documents,\ w\ \ ic\ \ include\ all\ elements\ of\ combo_{ij}^{s},\ labelled\ wit\ \ category\ wit\ \ index\ p\ in\ train\ set$
- $for\ j = 3\ and\ j = 7$: $\Sigma_{i3}^{1,2} = 2,: \Sigma_{i3}^{2,2} = 0, \Sigma_{i3}^{3,2} = 1, \Sigma_{i3}^{4,2} = 0$, $\Sigma_{i7}^{1,2} = 1, \Sigma_{i7}^{1,2} = 0, \Sigma_{i7}^{1,2} = 0, \Sigma_{i7}^{1,2} = 0$
- $\Lambda_{ij} := Label^{q}\ w\ \ ere\ q = \arg\max_{p} \Sigma_{ij}^{p}$: $\Lambda_{i1} =$ "SANAT", $\Lambda_{i2} =$ "DÜNYA", $\Lambda_{i3} =$ "DÜNYA" , Λ_{i4}="DÜNYA", Λ_{i5}="DÜNYA", $\Lambda_{i6} =$ "SPOR", Λ_{i7}="DÜNYA"
- Λ_{i}^{p}: counts of $\Lambda_{ij}\ w\ \ ic\ \ equals\ to\ Label^{p}$, $\Lambda_{i}^{1} = 5, \Lambda_{i}^{2} = 1,\ \Lambda_{i}^{3} = 1, \Lambda_{i}^{4} = 0$
- Π_{ij}^{p} for j = 1 and j = 5 : $\Pi_{i1}^{1,2} = 0, \Pi_{i1}^{2,2} = 0, \Pi_{i1}^{3,2} = 1, \Pi_{i1}^{4,2} = 0, \Pi_{i5}^{1,2} = 0.75, \Pi_{i5}^{2,2} = 0.25, \Pi_{i5}^{3,2} = 0, \Pi_{i5}^{4,2} = 0$

Application (computations)

- ρ_i^p: $\rho_i^1 = \frac{8}{7384} = 0.001$, $\rho_i^2 = \frac{3}{1568} = 0.002$, $\rho_i^3 = \frac{3}{229} = 0.013$, $\rho_i^4 = \frac{0}{116} = 0$
- $\overline{\Pi_i^p}$: $\overline{\Pi_i^1} = \frac{1+0.67+1+0.75+1}{4} = 0.884$, $\overline{\Pi_i^2} = \frac{0.25+1}{2} = 0.625$, $\overline{\Pi_i^3} = \frac{1+0.33}{2} = 0.665$, $\overline{\Pi_i^4} = 0$
- $\widehat{\Pi_i^p}$: $\widehat{\Pi_i^1} = 1$, $\widehat{\Pi_i^2} = 1$, $\widehat{\Pi_i^3} = 1$, $\widehat{\Pi_i^4} = 0$

Application (Predictions)

- $for\ i = 38296$
- $predict\_cb_1(doc_i, 2) = \text{SPOR}$
- $predict\_cb_2(doc_i, 2) = \text{SANAT}$
- $predict\_cb_3(doc_i, 2) = \text{DÜNYA}$
- $predict\_cb_4(doc_i, 2) = \text{DÜNYA}$
- $predict\_cb_5(doc_i, 2) = \text{SANAT}$

Application (Results)

		Confusion Matrix for Model 1 (count) s= 2						Confusion Matrix for Model 1 (percentage)					
		Prediction						Prediction (rounded to 2 digits)					
		DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total	DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total
Observed	DÜNYA	1204	47	519	47	25	1842	0.65	0.03	0.28	0.03	0.01	1
	SANAT	15	23	15	0	3	56	0.27	0.41	0.27	0	0.05	1
	SPOR	201	3	182	0	13	399	0.5	0.01	0.46	0	0.03	1
	Teknoloji	18	0	8	2	0	28	0.64	0	0.29	0.07	0	1
Accuracy Rate For Model 1													
0.61													

		Confusion Matrix for Model 2 (count) s= 2						Confusion Matrix for Model 2 (percentage)					
		Prediction						Prediction (rounded to 2 digits)					
		DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total	DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total
Observed	DÜNYA	1286	174	130	227	25	1842	0.7	0.09	0.07	0.12	0.01	1
	SANAT	3	41	9	0	3	56	0.05	0.73	0.16	0	0.05	1
	SPOR	30	18	324	14	13	399	0.08	0.05	0.81	0.04	0.03	1
	Teknoloji	14	0	2	12	0	28	0.5	0	0.07	0.43	0	1
Accuracy Rate For Model 2													
0.72													

Application (Results)

		Confusion Matrix for Model 3 (count) s= 2						Confusion Matrix for Model 3 (percentage)					
		Prediction						Prediction (rounded to 2 digits)					
		DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total	DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total
Observed	DÜNYA	1755	17	25	20	25	1842	0.95	0.01	0.01	0.01	0.01	1
	SANAT	39	8	6	0	3	56	0.7	0.14	0.11	0	0.05	1
	SPOR	159	8	214	5	13	399	0.4	0.02	0.54	0.01	0.03	1
	Teknoloji	28	0	0	0	0	28	1	0	0	0	0	1
Accuracy Rate For Model 3													
0.85													

		Confusion Matrix for Model 4 (count) s= 2						Confusion Matrix for Model 4 (percentage)					
		Prediction						Prediction (rounded to 2 digits)					
		DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total	DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total
Observed	DÜNYA	1805	3	7	2	25	1842	0.98	0	0	0	0.01	1
	SANAT	47	2	4	0	3	56	0.84	0.04	0.07	0	0.05	1
	SPOR	284	0	102	0	13	399	0.71	0	0.26	0	0.03	1
	Teknoloji	28	0	0	0	0	28	1	0	0	0	0	1
Accuracy Rate For Model 4													
0.82													

Application (Results)

		Confusion Matrix for Model 5 (count) s= 2						Confusion Matrix for Model 5 (percentage)					
		Prediction						Prediction (rounded to 2 digits)					
		DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total	DÜNYA	SANAT	SPOR	Teknoloji	No Prediction	Total
Observed	DÜNYA	977	289	189	362	25	1842	0.53	0.16	0.1	0.2	0.01	1
	SANAT	3	40	10	0	3	56	0.05	0.71	0.18	0	0.05	1
	SPOR	25	27	308	26	13	399	0.06	0.07	0.77	0.07	0.03	1
	Teknoloji	10	0	3	15	0	28	0.36	0	0.11	0.54	0	1
Accuracy Rate For Model 5													
0.58													

A Note:

All predictions of 25,2,13 documents labelled with «DÜNYA», «SANAT» and «SPOR» respectively are «No prediction». Because no combinations, with $s = 2$ stems,of those documents in test set are covered by a document in train set. Trivially prediction based combinations of stems of these documents, with $s > 2$ stems, are «No Predidiction».

End of Chapter Two