Irony Detection in Twitter with Imbalanced Scenarios

Corpora distribution

Figure 1 shows the distribution of ironic and non-ironic tweets in the used corpora.

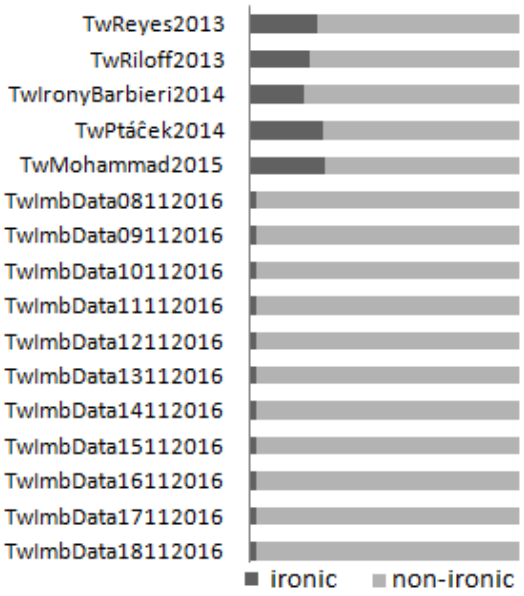

Figure 1: "Ironic" and "non-ironic" tweets distribution in the corpora.

Results

We experimented with a set of classifiers composed by: Naive Bayes (NB), Decision Tree (J48), Support Vector Machine (SVM), and Random Forest (RF). With the aim to compensate class imbalance distribution four techniques were applied: random oversampling (ROS), random undersampling (RUS), synthetic minority oversampling (SMOTE), and cost-sensitive learning (COST). We considered two evaluation metrics: the area under the ROC curve (AUC) and F-1 score. Tables 1-4 show the results obtained.

	F1					**AUC**				
Dataset	Original	COST	ROS	RUS	SMOTE	Original	COST	ROS	RUS	SMOTE
TwReyes2013	0.57	**0.589**	**0.581**	**0.601**	**0.67**	0.821	0.821	0.821	**0.837**	**0.86**
TwRiloff2013	0.487	**0.498**	**0.49**	**0.498**	0.467	0.742	0.742	**0.745**	**0.746**	0.724
TwIronyBarbieri2014	0.6	0.573	0.598	0.591	0.597	0.852	0.852	0.851	0.848	0.849
TwSarcasmBarbieri2014	0.702	0.652	0.686	0.689	0.671	0.908	0.908	0.908	**0.91**	0.886
TwPtáček2014	0.519	**0.543**	**0.534**	**0.531**	**0.563**	0.739	0.739	0.738	0.736	**0.766**
TwMohammad2015	0.412	**0.443**	**0.417**	**0.45**	**0.42**	0.635	0.635	0.631	**0.637**	0.619
TwImbData08112016	0.061	0.057	0.061	0.06	0.06	0.662	0.662	0.662	0.659	0.637
TwImbData09112016	0.094	0.076	0.089	0.089	0.08	0.781	0.781	**0.782**	0.779	0.815
TwImbData10112016	0.107	0.087	0.099	0.1	0.095	0.797	0.797	0.797	0.795	**0.831**
TwImbData11112016	0.105	0.088	0.101	0.101	0.098	0.799	0.799	0.799	**0.8**	**0.827**
TwImbData12112016	0.106	0.088	0.1	0.096	0.095	0.796	0.796	0.796	**0.788**	**0.817**
TwImbData13112016	0.101	0.086	0.095	0.094	0.083	0.788	0.788	0.788	0.786	**0.804**
TwImbData14112016	0.097	0.081	0.094	0.094	0.082	0.786	0.786	0.786	0.787	**0.8**
TwImbData15112016	0.097	0.084	0.096	0.095	0.082	0.789	0.789	0.789	0.783	**0.823**
TwImbData16112016	0.097	0.081	0.096	0.097	0.077	0.791	0.791	0.791	0.793	0.783
TwImbData17112016	0.101	0.085	0.097	0.094	0.088	0.792	0.792	0.792	0.788	**0.834**
TwImbData18112016	0.103	0.082	0.094	0.09	0.078	0.788	0.788	0.788	0.783	**0.793**

Table 1: Results obtained by using NB for original distribution as well as applying class imbalance treatment techniques in terms of F1 and AUC for each dataset.

	F1					**AUC**				
Dataset	Original	COST	ROS	RUS	SMOTE	Original	COST	ROS	RUS	SMOTE
TwReyes2013	0.826	0.806	0.815	0.803	0.815	0.888	**0.915**	0.861	**0.891**	0.881
TwRiloff2013	0.436	**0.516**	0.427	**0.438**	**0.46**	0.635	**0.735**	0.627	0.631	**0.65**
TwIronyBarbieri2014	0.796	0.724	0.766	0.722	0.767	0.879	**0.903**	0.851	0.862	0.862
TwSarcasmBarbieri2014	0.84	0.787	0.815	0.789	0.821	0.903	**0.934**	0.876	0.897	0.897
TwPtáček2014	0.584	**0.629**	0.578	**0.615**	**0.595**	0.709	**0.785**	0.685	**0.736**	**0.722**
TwMohammad2015	0.378	**0.416**	0.338	**0.413**	**0.397**	0.56	**0.57**	0.535	0.557	0.564
TwImbData08112016	0.009	**0.07**	**0.091**	**0.068**	**0.078**	0.526	**0.726**	**0.543**	**0.642**	**0.65**
TwImbData09112016	0.258	0.125	**0.266**	0.125	0.225	0.785	**0.863**	0.639	0.778	0.725
TwImbData10112016	0.332	0.15	0.327	0.148	0.272	0.793	**0.882**	0.675	**0.828**	0.722
TwImbData11112016	0.278	0.126	**0.295**	0.123	0.212	0.792	**0.861**	0.657	0.768	0.685
TwImbData12112016	0.283	0.133	**0.285**	0.134	0.255	0.74	**0.863**	0.65	**0.809**	**0.746**
TwImbData13112016	0.278	0.131	**0.288**	0.129	0.235	0.744	**0.87**	0.654	**0.788**	0.732
TwImbData14112016	0.27	0.122	**0.271**	0.121	0.218	0.783	**0.861**	0.645	**0.788**	0.729
TwImbData15112016	0.252	0.124	**0.258**	0.128	0.2	0.782	**0.853**	0.636	0.771	0.699
TwImbData16112016	0.287	0.126	0.271	0.129	0.236	0.767	**0.858**	0.645	**0.787**	0.699
TwImbData17112016	0.284	0.132	**0.292**	0.126	0.225	0.756	**0.872**	0.652	**0.792**	0.725
TwImbData18112016	0.237	0.123	**0.246**	0.124	0.211	0.78	**0.849**	0.63	0.769	0.716

Table 2: Results obtained by using DT for original distribution as well as applying class imbalance treatment techniques in terms of F1 and AUC for each dataset.

	F1					**AUC**				
Dataset	Original	COST	ROS	RUS	SMOTE	Original	COST	ROS	RUS	SMOTE
TwReyes2013	0.801	0.709	0.79	0.789	0.799	0.859	0.852	**0.881**	**0.881**	0.883
TwRiloff2013	0.192	**0.429**	**0.54**	**0.528**	**0.557**	0.548	**0.627**	**0.73**	**0.723**	**0.743**
TwIronyBarbieri2014	0.743	0.496	0.717	0.713	0.729	0.824	0.74	**0.844**	**0.843**	**0.848**
TwSarcasmBarbieri2014	0.815	0.66	0.785	0.777	0.797	0.878	0.859	**0.897**	**0.893**	**0.897**
TwPtáček2014	0.502	**0.56**	**0.602**	**0.601**	**0.604**	0.665	**0.71**	**0.745**	**0.744**	**0.745**
TwMohammad2015	0.035	**0.432**	**0.427**	**0.445**	**0.386**	0.506	0.5	**0.59**	**0.6**	**0.583**
TwImbData08112016	0	**0.038**	**0.077**	**0.076**	**0.075**	0.5	0.5	**0.671**	**0.671**	**0.665**
TwImbData09112016	0	**0.063**	**0.144**	**0.131**	**0.15**	0.5	**0.704**	**0.831**	**0.813**	**0.826**
TwImbData10112016	0	**0.064**	**0.18**	**0.16**	**0.21**	0.5	**0.705**	**0.856**	**0.838**	**0.844**
TwImbData11112016	0	**0.054**	**0.169**	**0.14**	**0.184**	0.5	**0.649**	**0.84**	**0.815**	**0.82**
TwImbData12112016	0	**0.062**	**0.175**	**0.148**	**0.19**	0.5	**0.696**	**0.848**	**0.814**	**0.837**
TwImbData13112016	0	**0.061**	**0.17**	**0.153**	**0.176**	0.5	**0.692**	**0.841**	**0.816**	**0.827**
TwImbData14112016	0	**0.063**	**0.16**	**0.143**	**0.174**	0.5	**0.7**	**0.843**	**0.814**	**0.823**
TwImbData15112016	0	**0.058**	**0.155**	**0.139**	**0.158**	0.5	**0.675**	**0.828**	**0.805**	**0.827**
TwImbData16112016	0	**0.062**	**0.162**	**0.141**	**0.176**	0.5	**0.697**	**0.834**	**0.814**	**0.827**
TwImbData17112016	0	**0.059**	**0.164**	**0.146**	**0.167**	0.5	**0.678**	**0.839**	**0.813**	**0.833**
TwImbData18112016	0	**0.062**	**0.151**	**0.139**	**0.158**	0.5	**0.697**	**0.834**	**0.811**	**0.811**

Table 3: Results obtained by using SVM for original distribution as well as applying class imbalance treatment techniques in terms of F1 and AUC for each dataset.

	F1					AUC				
Dataset	Original	COST	ROS	RUS	SMOTE	Original	COST	ROS	RUS	SMOTE
TwReyes2013	0.854	0.835	**0.858**	0.829	**0.859**	0.972	0.971	0.972	0.969	0.972
TwRiloff2013	0.329	**0.534**	**0.446**	**0.526**	**0.514**	0.797	**0.805**	**0.808**	0.79	**0.81**
TwIronyBarbieri2014	0.814	0.801	**0.821**	0.794	**0.822**	0.958	0.958	**0.96**	0.956	**0.959**
TwSarcasmBarbieri2014	0.865	0.841	**0.869**	0.836	**0.869**	0.975	0.975	**0.976**	0.972	0.975
TwPtáček2014	0.624	**0.669**	**0.659**	**0.671**	**0.679**	0.888	0.886	0.888	0.884	0.888
TwMohammad2015	0.134	**0.467**	**0.232**	**0.481**	**0.303**	0.661	0.661	**0.663**	**0.658**	**0.649**
TwImbData08112016	0.039	**0.092**	**0.048**	**0.082**	0.05	0.758	**0.791**	**0.777**	**0.783**	**0.783**
TwImbData09112016	0.051	**0.151**	**0.091**	**0.134**	**0.13**	0.903	**0.926**	**0.918**	**0.918**	**0.916**
TwImbData10112016	0.042	**0.166**	**0.095**	**0.145**	**0.165**	0.922	**0.935**	**0.936**	**0.924**	**0.926**
TwImbData11112016	0.046	**0.154**	**0.074**	**0.132**	**0.107**	0.881	**0.915**	**0.905**	**0.906**	**0.901**
TwImbData12112016	0.049	**0.158**	**0.096**	**0.133**	**0.134**	0.902	**0.923**	**0.917**	**0.907**	**0.907**
TwImbData13112016	0.069	**0.155**	**0.087**	**0.132**	**0.117**	0.895	**0.925**	**0.918**	**0.913**	**0.91**
TwImbData14112016	0.075	**0.145**	**0.105**	**0.132**	**0.115**	0.901	**0.919**	**0.922**	**0.909**	**0.907**
TwImbData15112016	0.05	**0.141**	**0.072**	**0.125**	**0.096**	0.891	**0.909**	**0.908**	**0.899**	**0.91**
TwImbData16112016	0.057	**0.149**	**0.086**	**0.129**	0.088	0.904	**0.917**	**0.918**	**0.907**	**0.911**
TwImbData17112016	0.028	**0.148**	**0.056**	**0.134**	**0.098**	0.894	**0.915**	**0.912**	**0.906**	**0.914**
TwImbData18112016	0.027	**0.141**	**0.05**	**0.129**	**0.071**	0.884	**0.914**	**0.903**	**0.902**	**0.895**

Table 4: Results obtained by using RF for original distribution as well as applying class imbalance treatment techniques in terms of F1 and AUC for each dataset.

Critical statistical difference diagrams

Figure 2 shows all the pairwise comparisons of the learning algorithms in the original dataset distribution.

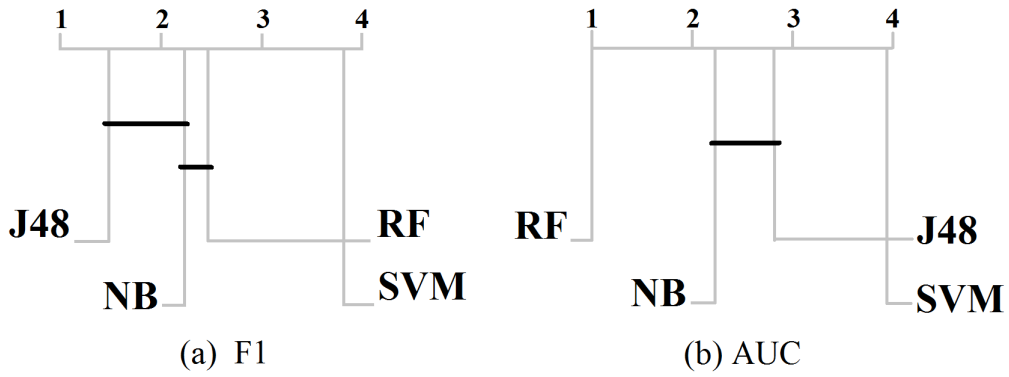

Figure 2: All pairwise comparison of algorithms using the original class distribution

Figures 3, 4, 5 and 6 show the critical differences diagrams for applying the class imbalance treatment techniques to NB, J48, SVM and RF, respectively.

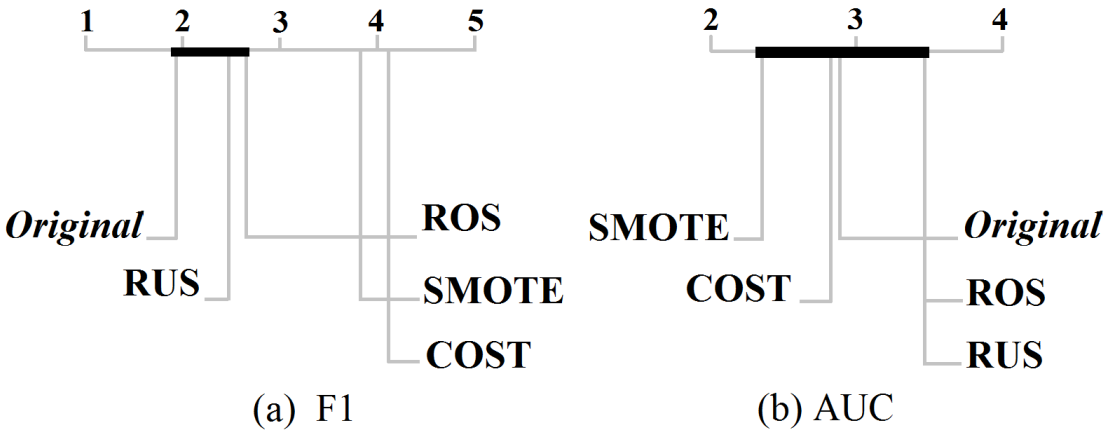

Figure 3: Comparison for the NB algorithm.

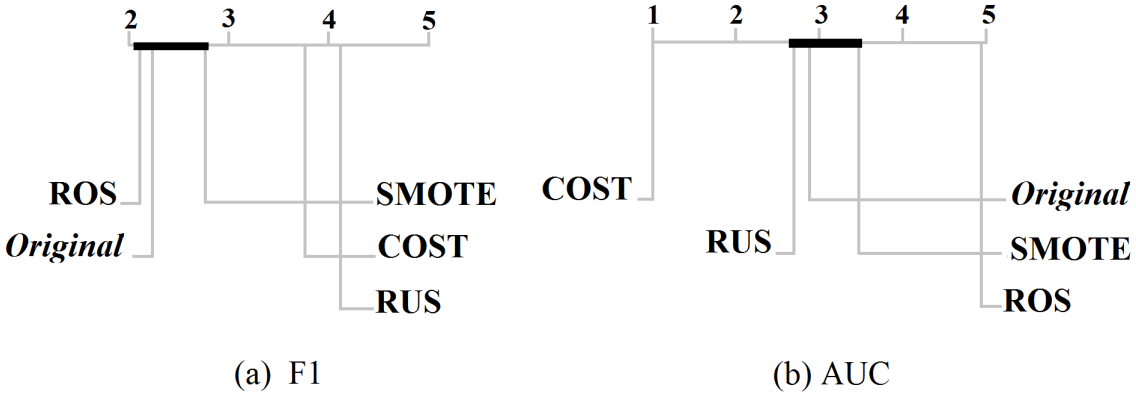

Figure 4: Comparison for the J48 algorithm.

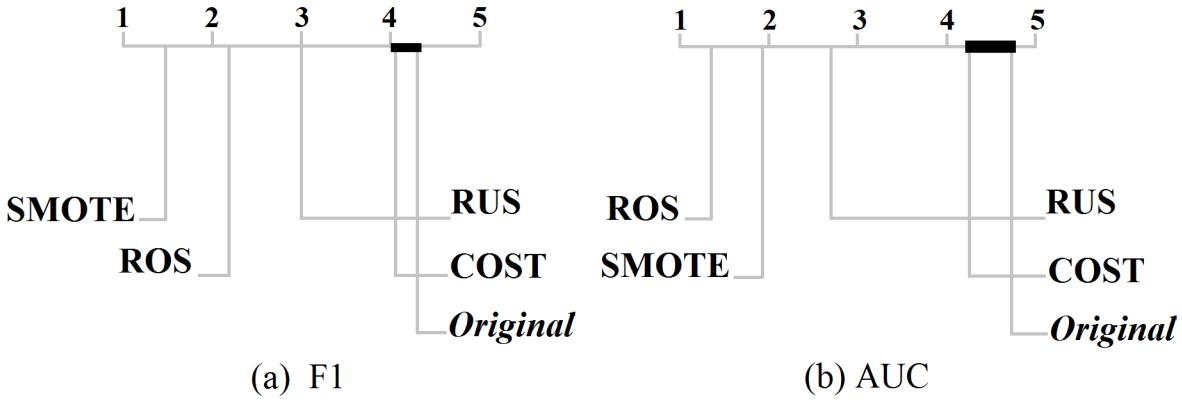

Figure 5: Comparison for the SVM algorithm.

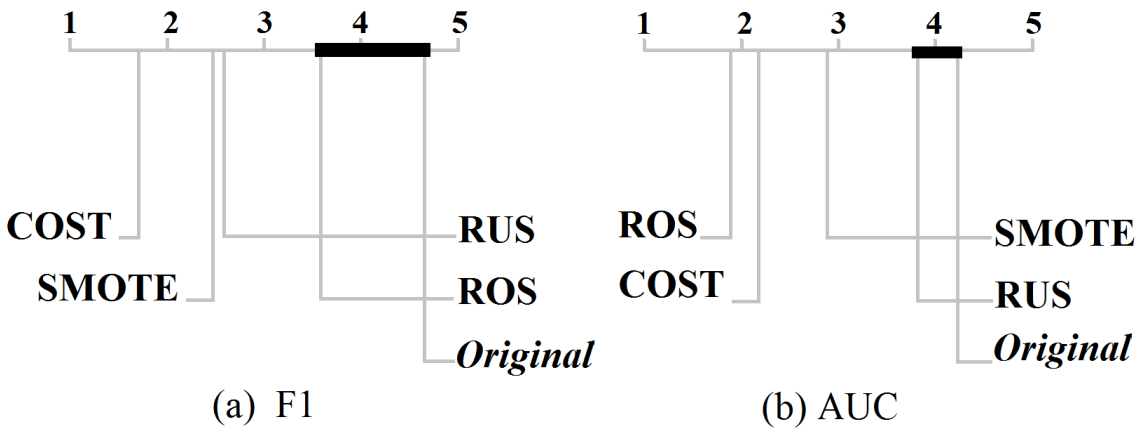

Figure 6: Comparison for the RF algorithm.