UNIVERSITÀ DEGLI STUDI DI TRENTO

Department of Engineering and Information Science

Master in Computer Science

Research Project in Multimedia Data Security

Classification of Sharing Applications

December 5, 2019

Supervisors:
Prof. Giulia Boato
PHD. Quoc Tin Phan

Student:
Kritjan Gjika

Academic Year 2018/2019

Contents

1 Double Scenario Classification of the last shared app, KFold Validation

Starting with fitting randomly the classifiers, there are some statistics of the data used for the first test:

	count train	count test
messenger	302	748
telegram	314	736
whatsapp	340	710
original	94	256

1.1 Logistic regression results:

Confusion matrix with number of sample and with normalization:

	messenger	telegram	whatsapp	original
messenger	746	0	2	0
telegram	0	618	118	0
whatsapp	0	216	494	0
original	0	0	8	248

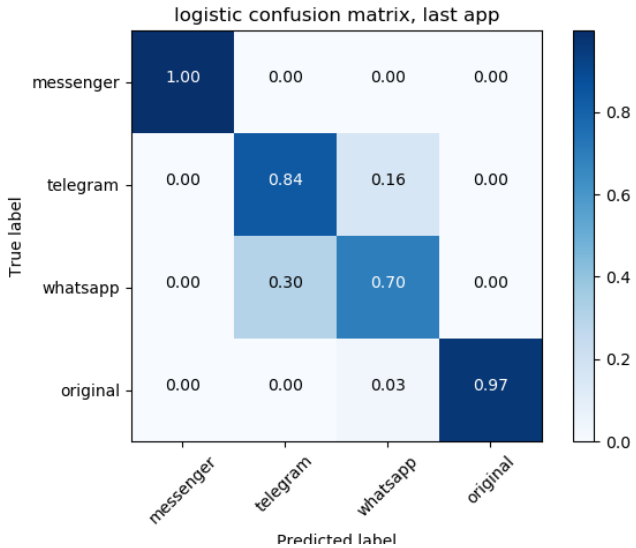

Figure 1.1: logistic regression, last app classified

Result of the KFold validation with 10 bins:

2 Double Scenario Classification of the first and last shared app, KFold Validation

Starting with fitting randomly the classifiers, there are some statistics of the data used for the first test:

	count train	count test
mess_mess	96	254
tele_mess	99	251
what_mess	107	243
mess_tele	98	252
tele_tele	111	239
what_tele	105	245
mess_what	116	234
tele_what	103	247
what_what	121	229
original	94	256

2.1 Logistic regression results:

Confusion matrix with number of sample and with normalization:

	m_m	m_t	m_w	t_m	t_t	t_w	w_m	w_t	w_w	original
mess_mess	248	5	1	0	0	0	0	0	0	0
tele_mess	2	233	14	0	0	0	2	0	0	0
what_mess	13	26	202	0	0	2	0	0	0	0
mess_tele	0	0	0	65	116	3	0	68	0	0
tele_tele	0	0	0	66	57	2	0	114	0	0
what_tele	0	0	0	3	4	235	1	2	0	0
mess_what	0	0	0	1	0	1	80	0	152	0
tele_what	0	0	0	65	129	3	0	50	0	0
what_what	0	0	0	0	0	0	104	0	123	2
original	0	0	0	0	0	0	0	0	5	251

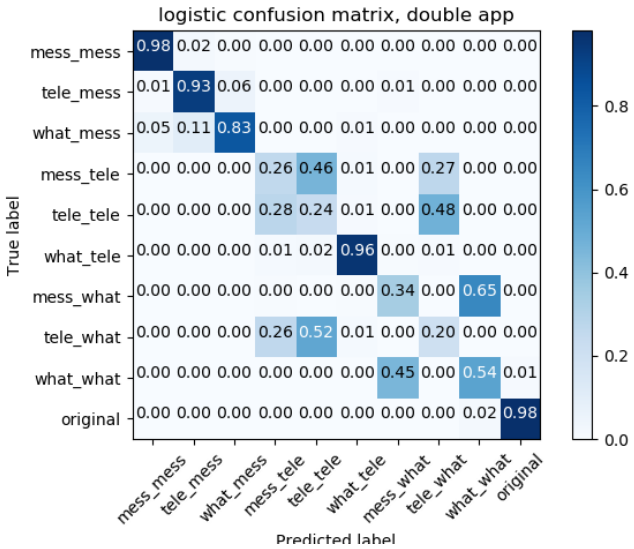

Figure 2.1: logistic regression, last app classified

Result of the KFold validation with 10 bins:

0.6000	0.5619	0.6381	0.6095	0.6190	0.6667	0.6000	0.5810	0.5429	0.6190

The mean is : 0.603810

2.2 Linear Support Vector Machine results:

Confusion matrix with number of sample and with normalization:

	m_m	m_t	m_w	t_m	t_t	t_w	w_m	w_t	w_w	original
mess_mess	244	5	2	0	2	0	0	0	1	0
tele_mess	1	217	14	3	2	0	10	1	3	0
what_mess	14	17	192	3	0	1	9	4	3	0
mess_tele	0	0	0	68	106	5	1	71	1	0
tele_tele	0	1	0	64	49	4	0	121	0	0
what_tele	0	0	0	4	4	234	1	2	0	0
mess_what	0	0	0	2	1	0	83	0	148	0
tele_what	0	1	0	61	131	5	0	49	0	0
what_what	0	0	0	1	1	0	101	1	125	0
original	0	0	0	2	1	0	0	0	5	248

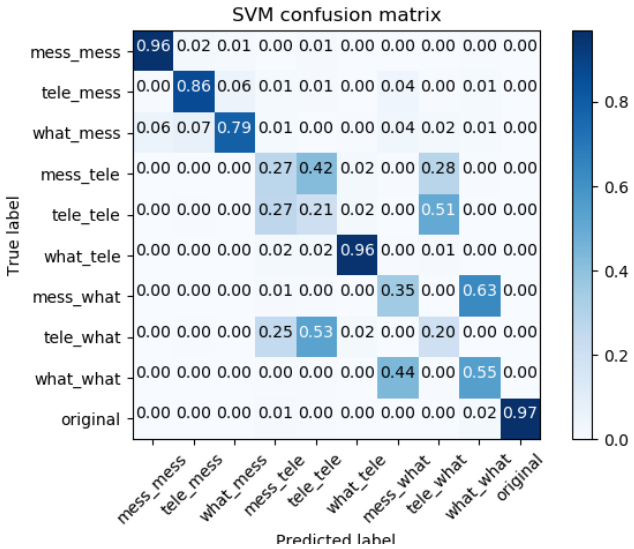

Figure 2.2: linear SVM, last app classified

Result of the KFold validation with 10 bins:

0.6000	0.5429	0.6095	0.5810	0.6000	0.7048	0.6381	0.6000	0.5810	0.6190

The mean is : 0.607619

2.3 Random forest results:

Confusion matrix with number of sample and with normalization:

	m_m	m_t	m_w	t_m	t_t	t_w	w_m	w_t	w_w	original
mess_mess	236	7	9	0	0	0	0	0	1	1
tele_mess	13	214	24	0	0	0	0	0	0	0
what_mess	24	21	188	0	0	0	2	0	5	3
mess_tele	0	0	0	32	116	3	0	101	0	0
tele_tele	0	0	0	72	46	4	0	117	0	0
what_tele	0	0	0	1	2	239	0	3	0	0
mess_what	1	0	0	0	0	0	70	0	163	0
tele_what	0	0	0	71	128	5	0	43	0	0
what_what	0	0	0	0	0	0	127	0	98	4
original	0	0	0	0	0	1	0	0	1	254

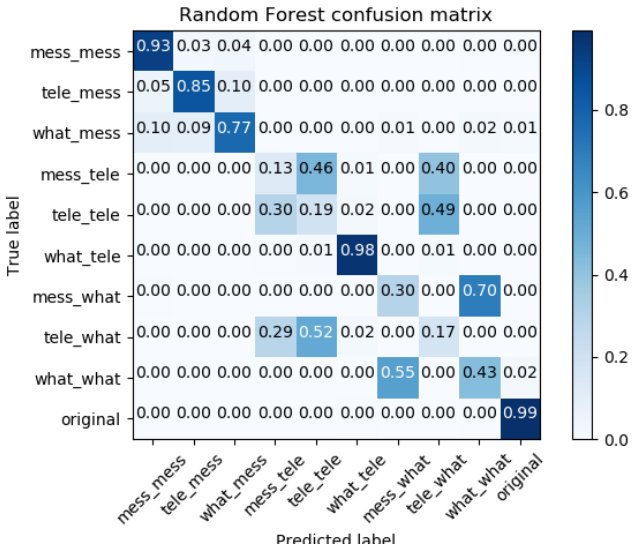

Figure 2.3: random forest, last app classified

Result of the KFold validation with 10 bins:

0.5333	0.5048	0.6095	0.5238	0.5810	0.5905	0.5714	0.6286	0.5524	0.5619

The mean is : 0.565714

3 Single and double scenario, KFold Validation

Starting with fitting randomly the classifiers, there are some statistics of the data used for the first test:

	count train	count test
mess	97	253
tele	335	715
what	219	481
mess_mess	99	251
tele_mess	113	237
what_mess	93	257
mess_tele	89	261
what_tele	103	247
mess_what	106	244
original	111	239

3.1 Logistic regression results:

Confusion matrix with number of sample and with normalization:

	m	t	w	m_m	m_t	m_w	t_m	t_w	w_m	original
mess	184	0	0	46	4	19	0	0	0	0
tele	0	710	0	0	0	0	0	5	0	0
what	0	0	419	0	0	0	0	0	58	4
mess_mess	21	0	0	220	6	4	0	0	0	0
tele_mess	0	0	0	0	215	22	0	0	0	0
what_mess	0	0	0	0	88	169	0	0	0	0
mess_tele	0	256	0	0	0	0	0	5	0	0
what_tele	0	9	0	0	0	0	2	236	0	0
mess_what	0	0	191	0	0	0	0	1	52	0
original	0	0	0	0	0	0	0	0	0	239

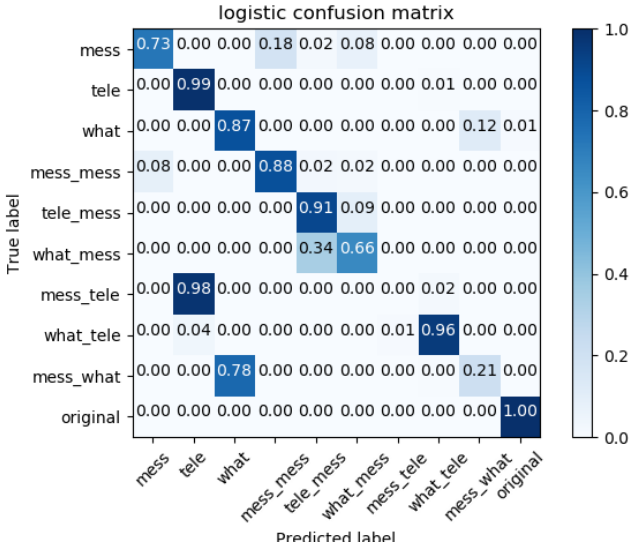

Figure 3.1: logistic regression, last app classified

Result of the KFold validation with 10 bins:

0.8029	0.7810	0.7810	0.7737	0.7372	0.7353	0.7059	0.8162	0.7721	0.7794

The mean is : 0.768474

3.2 Linear Support Vector Machine results:

Confusion matrix with number of sample and with normalization:

	m	t	w	m_m	m_t	m_w	t_m	t_w	w_m	original
mess	177	0	0	49	4	23	0	0	0	0
tele	0	715	0	0	0	0	0	0	0	0
what	0	0	392	0	0	0	0	0	85	4
mess_mess	21	0	0	222	6	2	0	0	0	0
tele_mess	0	0	0	0	212	25	0	0	0	0
what_mess	0	0	0	2	42	213	0	0	0	0
mess_tele	0	261	0	0	0	0	0	0	0	0
what_tele	0	5	0	0	0	0	0	242	0	0
mess_what	0	0	161	0	0	0	0	1	82	0
original	0	0	0	0	0	0	0	0	0	239

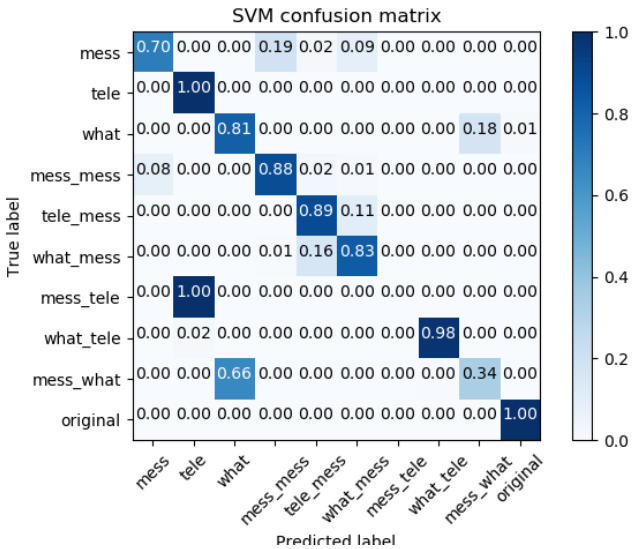

Figure 3.2: linear SVM, last app classified

Result of the KFold validation with 10 bins:

0.8248	0.7883	0.7883	0.8102	0.7518	0.7868	0.7279	0.7868	0.7941	0.8015

The mean is : 0.786056

3.3 Random forest results:

Confusion matrix with number of sample and with normalization:

	m	t	w	m_m	m_t	m_w	t_m	t_w	w_m	original
mess	186	0	1	42	2	19	0	1	2	0
tele	0	597	0	0	0	0	115	3	0	0
what	0	0	354	0	0	0	0	0	123	4
mess_mess	48	0	0	196	3	3	0	0	1	0
tele_mess	0	0	0	3	219	15	0	0	0	0
what_mess	13	0	7	15	22	198	0	0	2	0
mess_tele	0	195	0	0	0	0	65	1	0	0
what_tele	0	4	0	0	0	0	1	242	0	0
mess_what	1	0	184	0	0	0	0	0	59	0
original	2	0	5	1	0	0	0	1	0	230

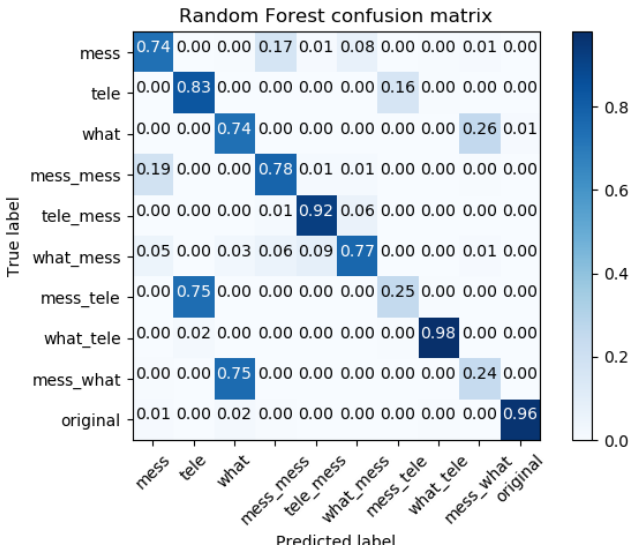

Figure 3.3: random forest, last app classified

Result of the KFold validation with 10 bins:

0.7518	0.7664	0.7007	0.7445	0.6642	0.8015	0.7132	0.7206	0.6544	0.7353

The mean is : 0.725274

Bibliography