- Page 1 -

Binary Text Classification

Enron Email Dataset			
# of train docs	4138	94	187
Original	**0.9500**	**0.5500**	**0.9000**
back-translation			0.8500
thesarius			0.9000
thesaurus geom.*			1.0

*thesaurus with geometrical distribution

With tfidf:

Multiclass Text Classification

20 Newsgoups (use 4 categories of 20 - 'alt.atheism', 'comp.graphics', 'sci.med', 'talk.religion.misc')						
# of train docs		2035	100	200	300	2000
Original	**NB**	**0,8295**	**0,4317**	**0,5771**	**0,7299**	**0,8288**
	SVM	**0,8657**	**0,6701**	**0,7247**	**0,7956**	**0,8605**
	Log Regr		**0,6605**	**0,7107**	**0,7624**	**0,8214**
Back Translation	NB			0,5137	0,5616 *	x
	SVM			0,6856	0,6952 *	x
	Log Regr			0,6679	0,6524 *	x
Thesarius	NB			0,5321		0,8384
	SVM			0,7077		0,8443
	Log Regr			0,6731		0,8066
Thesarius Geom. Distribution	NB			0,5528		0,8502
	SVM			0,7026		0,8517
	Log Regr			0,6686		0,8162
Back Tr + Thesarius	NB				0,5579	x
	SVM				0,6967	x
	Log Regr				0,6590	x
				from 100 original	from 100 original	from 1000 original

** (de+fr) together*

x Requst error after about 130 samples (google.translate)
<u>Goofle Translater API limitations:</u>
The total of all texts to be translated must not exceed 10000 characters.
The maximum number of array elements is 2000.
On 20 newsgroup Dataset - not more than 130 train documents at one time.

---- > Page 2

- Page 2 -

Improved Pipeline: Classification with BOW, TFIDF, word2vec and TFIDF word2vec:

Multiclass Text Classification

20 Newsgoups (use 4 categories of 20 - 'alt.atheism', 'comp.graphics', 'sci.med', 'talk.religion.misc')						
# of train docs		2035	100	200	300	2000
Original	**NB with BOW**		**0,7203**	**0,7314**	**0,7815**	**0,8915**
	SVM with BOW		**0,5838**	**0,5897**	**0,6546**	**0,8155**
	NB with tfidf		**0,4317**	**0,5771**	**0,7299**	**0,8288**
	SVM with tfidf		**0,6937**	**0,7306**	**0,7838**	**0,8738**
	SVM with averaged word2vec		**0,3292**	**0,3727**	**0,3144**	**0,6576**
	SVM with tfidf weighted averaged word2vec		**0,2856**	**0,3402**	**0,4502**	**0,6613**
Back Translation				*from 100 original*	*from 100 original **	*from 1000 original*
	NB with BOW			0,7351	0,7424	x
	SVM with BOW			0,5786	0,5956	x
	NB with tfidf			0,5137	0,5616	
	SVM with tfidf			0,6915	0,6937	
	SVM with averaged word2vec			0,3653	0,3720	
	SVM with tfidf weighted averaged word2vec			0,3321	0,4022	x
Thesarius				*from 100 original*	*from 100 original*	*from 1000 original*
	NB with BOW			0,7240		0,8723
	SVM with BOW			0,6111		0,7875
	NB with tfidf			0,5321		0,8384
	SVM with tfidf			0,6923		0,8517
	SVM with averaged word2vec			0,2878		0,6531
	SVM with tfidf weighted averaged word2vec			0,3159		0,6524

Back Tr + Thesarius				*from 100 original*	*from 100 original*	*from 1000 original*
	NB with BOW				0,7395	x
	SVM with BOW				0,6081	x
	NB with tfidf				0,5579	
	SVM with tfidf				0,6959	
	SVM with averaged word2vec				0,3232	
	SVM with tfidf weighted averaged word2vec				0,3284	x