ARABIC SENTIMENT ANALYSIS ON TWITTER

Azouz saoussen [1] Haffar Nafaa [2] Zrigui Mounir [2] [1]Polytech University, [2]the faculty of science of Monastir

Introduction

My study involves providing the current state of the art in Arabic sentiment analysis as well as testing the most commonly used approach in the ASA which is the ML approch.The majority of studies found that the *SVM* and *NB* were the two most efficient classifiers in the ML technique. A new NLP toolkit has been developed to aid in the pre-processing step of ASA. My purpose was to conduct my own comparition study on eight ML classifers and to test this toolkit..

State of the Art ML approach

Purpose	Accuracy	Used algorithms	References
tackling the DA and SA for MSA	60.32%	SVM	Al-Subaihin and Al-Khalif 2014
SA on Arabizi	86.9%	SVM and NB	Duwairi et al. 2016
SA for Arabic tweets	84.62%	SVM, NB, TF-IDF and BTO	Al-Rubaiee et al.2016
Develop Arabic Senti lexicon for SA and create a Corpus (MASC)	Between 40% and 68.7%	SVM, KNN	Al-Moslmi et al. 2017

Data collection and pre-processing

I used a dataset with 55k tweet Ids. Then the tweets were extracted using Twitter API and four python modules (Tweepy, tqdm, pandas and pickle) And then for cleaning data, i created a model based on Farasapy, here are the steps:

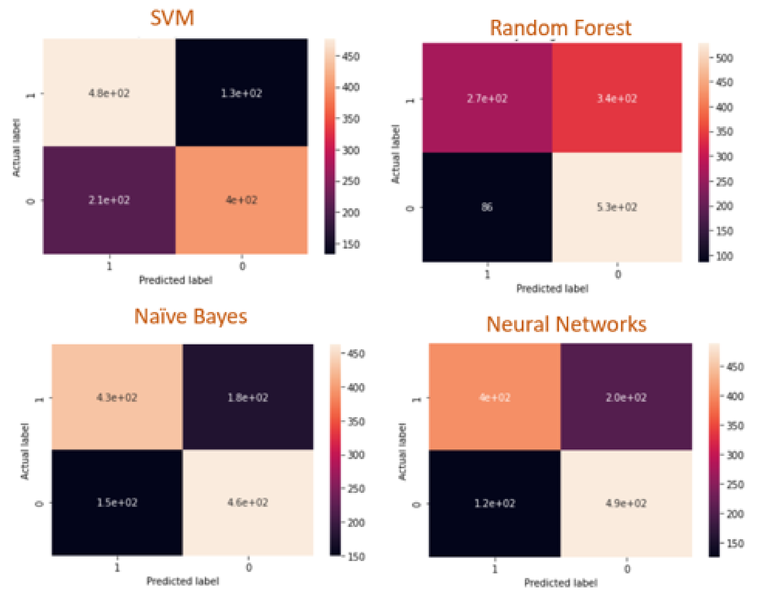
- remove emails
- remove underscore
- remove phone numbers
- remove all punctuations
- remove non arabic words
- remove URLs
- remove extra spaces
- strip tashkeel and tatweel

	Tweet_id	normalized_text	sentiment	text
0	1143297865445457921	مباراه اوروجواي وتشيلي بث مباشر كوبا امريكا حص...	Neutral	مباراه اوروجواي وتشيلي بث مباشر كوبا امريكا 2\n\nمباشر...
1	1221882773272657920	ربي هذا حنين يرهقني اجمعني اب في فردوس	Negative	ربّـاه ! هذا الحنين يُرهِقني اجمعني بأبـ\n\n...
2	1143248282703060994	طايف قام حذف اسم مسيء من برنامج تمبر بوك تواصل...	Neutral	#الطايف بحذف الأسماء\n\nقوم من برنامج\n\nالمسيئة ...
3	1242908498427592705	شخص مصاب فيروس انفلونزا لا اصاب فيروس كورونا م...	Neutral	الشخص المصاب بفيروس الانفلونزا لا يصاب بفيروس ...
4	1145815687065198601	كان خايفين احد يترجم كمنشي جو بقر قدمولهم خبر ب...	Negative	كنا خايفين احد يترجم للكمنشيات وجو البقر قدمول...

Model and features

My model consists of combination of multiple classifiers. And for features i used *TfidfVectorizer* from sklearn converts a collection of raw documents to a matrix of TF-IDF features. Below, some of classifiers that i used and their confusion matrices:

Results & Conclusions



Algorithm	Accuracy(%)
LogisticRegression	77,8
Naïve Bayes	80,1
Linear SVM	78,1
RBF SVM	77,7
Random Forest	76
MLP	79,5
AdaBoost	72
GradientBoosting	72

azzouz.saoussen2712@gmail.com