Tunisian Dialect and Modern Standard Arabic Dataset for Sentiment Analysis : Tunisian Election Context.

Karim Sayadi[1 2], Marcus Liwicki[3], Rolf Ingold[3], Marc Bui[2]

[1] Sorbonne University, UPMC, Univ Paris 06
karim.sayadi@upmc.fr
[2] EPHE, CHArt Laboratory EA 4004,
4-14 Rue Ferrus, 75014 Paris, France.
[3] DIVA Group, University of Fribourg,
Bd. de Prolles 90, 1700 Fribourg, Switzerland.

Abstract. With the growth in use of social media platform, Sentiment Analysis methods become more and more popular while for classical Arabic there are many datasets available, most of the Arab countries' Dialects suffer from very limited resources for training an automated system. In this article, we study Sentiment Analysis applied on the Tunisian Dialect and Modern Standard Arabic by providing a manually annotated dataset. On this dataset, we performed feature selection and trained 6 classifiers to provide a benchmark for future studies.

1 Introduction

Social media web sites (e.g. Twitter, Facebook, Tumblr) continue growing in popularity and are getting more and more notoriety among politics since the US presidential elections of 2008 [1], where analysts believe that the social media strategy of the Obama campaign contributed to his win. Since then an increasing amount of research [2] was performed on analyzing political content on social media [3]. Although it has a remarkable potential for the political parties[4], to the best of the authors' knowledge there is no work on Arab elections in the research so far.

In this work, we make the first investigations of automated Sentiment Analysis for Tunisian Dialect and therefore provide an annotated dataset for the sentiment for developing and training automated Sentiment Analysis systems with a special focus on the elections context. We have chosen the micro-blogging service Twitter as our data source because it is one of the major sources of online political commentary and discussion in Tunisia and the access to the data is made easier than other social media web sites. In 2011 Twitter was a vital communication platform in Tunisia for the upraising against the dictatorship,

[4] 5.8 million active Arab users on Twitter. Recorded on March 2014 according to the http://www.arabsocialmediareport.com/News/description.aspx?NewsID=16

because it was fast to broadcast a message and more easy to gain visibility than other social media platforms.

Our data set is composed of a collection of short texts, of 140 characters at maximum, called tweets. The tweets were gathered during the period of the national assembly election and the presidential election. We started collecting data from Twitter with the beginning of the campaigns of the different political parties in October, 4th, 2014. These elections were important because they were the first elections after the adoption of the new constitution in January 2014. We stopped collecting the data in December, 23rd, 2014, just after the presidential election. During this period, people have used Twitter to comment the political speeches on TV or the news articles in written media. We believe that applying the Sentiment Analysis algorithms could provide to the political parties and the public opinion precious information about the election context.

1.1 Contributions

The contributions of this paper are as follows.

1. We present a manually annotated dataset composed of Tunisian Dialect (TN) and Modern Standard Arabic (MSA) for the Sentiment Analysis task.
2. We compare the performance of six different classifiers with the use of Information Gain feature selection method applied to Sentiment Analysis. This part can be considered as a benchmark experiments for future comparisons.
3. We investigate the statistics of the proposed dataset and the difference between trained classifiers on TN and MSA.

1.2 Outline

This paper is organized as follows : Section 2 provides an overview of the state of the state of art. The collection of the dataset and the related statistics are presented in Section 3. Section 4 presents the experimentations and investigation about the difference between trained classifiers on TN, MSA and the both. We conclude in Section 5 with a discussion and future works.

2 Related work

The micro-blogging service Twitter has become a central site where people can create and exchange content with a large audience. In the context of an election for example, people tend to use Twitter to express their opinions [10] and views about the political parties or candidates [11]. Emerging events or news are often followed almost instantly by a burst in Twitter volume [12], providing a unique opportunity to gauge the opinions and sentiments towards the electoral events.

We observe an increasing interest in the Arabic NLP community for opinion mining and sentiment analysis [13–17]. Nevertheless, due to the diversity of Arabic language and its different Dialects, there is a need to gather more data from

Table 1. Arabic Sentiment Analysis Datasets ordered by the size.

Cite/Year	Name	Size	MSA/Dialect	Source
[4]/2013	LSABR	63,257	MSA/Not Mentioned	GoodReads.com
[5]/2015	HTL	15,572	MSA/Not Mentioned	TripAdvisor.com
[5]/2015	RES	10,970	MSA/Not Mentioned	Qaym.com
[6]/2015	ASTD	10,000	MSA+Egyptian	Twitter.com
[7]/2014	ATC	8,868	MSA+JO	Twitter.com
[5]/2015	PROD	4,272	MSA/Not Mentioned	Souq.com
[8]/2014	MONTADA	3,097	MSA/LEV/EGY	Forums
[8]/2014	TGRD	3,015	MSA/not precised	Twitter.com
[9]/2014	THRIR	3,008	MSA/LEV/EGY	Wikipedia TalkPages
[5]/2015	MOV	1,524	MSA/Not Mentioned	Elcinemas.com

different countries to develop more generic tools. For example a word segmentation[18] or a morphological analysis tool [19] developed for MSA and Egyptian (EGY) produces non accurate results on Tunisian Dialect(TN) because it is very different from the MSA or EGY. Please refer to the analysis presented in this work [20].

In Table 1 we list the datasets that were collected for the Sentiment Analysis task. Three of them [6–8] are collected from Twitter and annotated. The work in [6] conducted an experimentation with 4 classifiers and did not distinguish between the EGY and MSA in the training process. The work in[7] collected MSA and Jordanian tweets and presented some statistics about the dataset. Finally, in [8] the authors studied the lexicon of their collected tweets and presented an analysis of the arabic subjective sentiments.

3 Dataset

3.1 Dataset collection and annotation

We collected almost 50,000 tweets from 8293 users with the "Twitter Streaming API". The tweets were published on Twitter's public message board between October 1st 2014 and 23 December 2014. The first date is prior to the election of the 217 seat National Tunisian Assembly and the second date is posterior to the presidential elections. After separating the Arabic from non Arabic text and further processing we obtained 10,000 tweets written in Arabic letters. The constitution of the corpus is based on several keywords that have been manually tested in Twitter.com search bar and discussed with different members of the Twitter community. In addition to the keywords that are in Arabic we added the most used hashtags by the communities as input for the research of the tweets (see Table 2). Hashtags help the users to get more visibility for their tweets and represent a sort of hub where all the tweets that are around the same subjects could easily be found.

Table 2. Examples of the used inputs, keywords and hashtags, for the Twitter crawlers.

English Translation	Arabic
Presidential elections	الَانتخَابَات الرِئَاسِيّة
Legislative elections	الَانتخَابَات التشريعية
Tunisian elections	الَانتخَابَات التُونِسِيّة
Two of the main hashtags used by the Twitter community in Tunisia to talk about the elections	#tnelec #tnprez

Table 3. Example of annotated tweets.

	Label	Definition	Example	English Translation
TN	**positve**	With a positive used indicator	برَافُو نورَا.	Well done ! Noura.
	negative	With a negative used indicator	مَاينجمش يحكُم وحدو و لو حب.	He cannot govern alone even if he wanted to.
	neutral	No emotions' indicators used	كيفَاش التوَانسة المقيمين بَالخَارج ينتَخبو ؟	How Tunisians living abroad vote ?
MSA	**positve**	//	بعثة تصف الَانتخَابَات التونسية بأَنهَا تعدُدِية و شفَافة.	An official representation describes the elections as diverse and transparent.
	negative	//	مرَاقبو الريَاسية يتذمرون من عدم وصول بطقَاتهم	Presidential election' watchers complain about not receiving their cards.
	neutral	//	التفَاصيل في مقَال الجزيرة	The details are in AlJazir' article.

We manually annotated 5514 examples taking into account clear indicators of positive or negative. If these indicators were absent we consider the tweets as neutral (please refer to Table 3). In case where annotators are unable to decide about a sentences they just delete it. We don't take into account in annotation process the semantic or the hidden meaning. All of the tweets that contained advertisement or redundancy were also deleted.

Table 4. Statistics about the content of the collected Dataset

Nb. Tweets	5514
Nb. words	49940
Max words / Tweet	27
Min words / Tweet	1
Av. words / Tweet	10
Nb. vocabularies	10553

3.2 Data set statistics

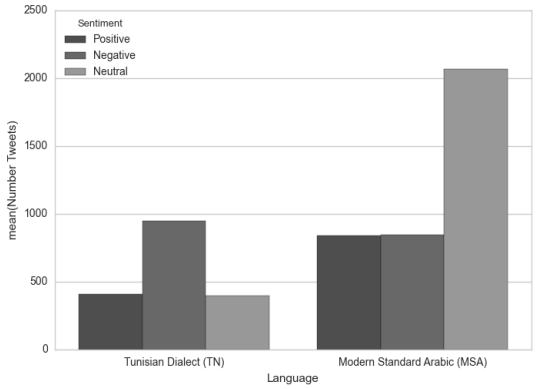


Fig. 1. Number of tweets for each category

The dataset has 5514 manually annotated tweets, 3760 of them are in MSA and 1754 are in TN. The proportion of the different classes is illustrated in Figure 3.2. For the MSA tweets we have a high number of neutral tweets this is due to the fact that the common language for the news is MSA. For the TN tweets we have a high number of negative tweets. The positive and neutral tweets are almost similar in numbers. A further analysis of reasons for this issues is beyond the scope of this paper. In Table 4 we present some statistics about the dataset.

4 Experimentations

4.1 Classifiers and Feature Selection

In this subsection, we briefly present the different classifiers as well as the feature selection method that we used in our experimentations.

Naive Bayes (NB) The NB classifier is the Bayes' rule (1) applied to documents (e.g tweets) and classes (e.g positive or negative). For a document d and a class c the Bayes' rule is written as following :

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{1}$$

$$P(c, x_1, ..., x_n) = P(c)P(x_1|c)...P(x_n|c) \tag{2}$$

Equation (2) represent the conditional independence where c is a binary class that include positive or negative and x is the feature (e.g a word in a tweet). The feature probabilities $P(x_i|c_j)$ are independent given the class c.

Support Vector Machines (SVM) SVM is a technique that builds an optimal separating hyperplane to find linear boundaries in the input feature space and separates two classes. SVM is a supervised binary classifier but we can used with multiple classes by classifying one class at a time.

k-Nearest Neighbor (NN) NN uses a majority vote to classify an object into the class of its k-nearest neighbor.In our experiments we set k=1.

Decision Trees A decision tree is a graph that has one root and a set of leafs. It takes as an input a description of a situation with it properties and output a decision yes or no. We used in our experimentation two methods implemented in Weka 3.6[5]. The first one is called **J48** and it is the Java implementation of the algorithm C4.5 [21]. The second one is called **PART**, it builds a partial C4.5 decision tree and generates a decision list.

Random Forest (RF) The term Random Forest was first introduced by Leo Breiman in 2001 [22], it can be considered as a set of randomly trained decision trees combined with a bagging method that reduce overfitting.

Table 5. Feature extraction on different grams and the selection ratio.

MSA/TN			**Nb. Grams**	**Nb. Grams IG**	**Ratio**
MSA+TN	$IG > 0.0$	1g	10553	406	3.85%
		1g+2g	33748	724	2.14%
		1g+2g+3g	58110	940	1.62%
TN	$IG > 0.0$	1g	7372	46	0.62%
		1g+2g	21803	56	0.25%
		1g+2g+3g	35862	57	0.15%

[5] weka.sourceforce.net

Feature extraction and selection We used Information Gain (IG) to extract the features that we used with the different classifiers. IG is a method used to extract the most prominent features with respect to class attribute. IG computes the expected reduction in entropy or the reduction in the uncertainty. When the entropy decreases the expected information increases. It is given by :

$$Info(S) = -\sum_{i=1} m(P_i) log_2(P_i) \tag{3}$$

where S denotes the set of instances, P is the probability that a random instance belongs to the class i and m is the number of classes (e.g 3 classes : positive, negative, neutral). Finally, to compute the IG we need to measure the number of bits required to encode the information of the classification of an instance in S by a feature F. This amount is given by :

$$Info_F(S) = -\sum_{j=1}^{v} \frac{|S_j|}{S} \times Info(S_j) \tag{4}$$

$$IG(F) = Info(S) - Info_F(S) \tag{5}$$

Where v is the number of partitions and $\frac{|S_j|}{S}$ is the weight of the partition(class) j and $Info(S_j)$ is the entropy of S_j computed with equation (1). The ranked features with an IG above 0.0 are selected.

In Table 5, we compare two features extraction. The first one is applied on the 5514 tweets composing our dataset. We have obtained 10553 unique grams separated with a space in each tweet. After the feature extraction we selected 406 unique grams with an IG above 0.0, for example "نحبها" in Tunisian which means "I love her". We repeated the processes for one and two grams features, like for example in Tunisian "اتعس من" which means "worse than". The second part of the Table is applied on only Tunisian Dialect. The ratio in the table represents the aggressivity of the feature selection, the smaller it is the more aggressive the feature selection is.

4.2 Results

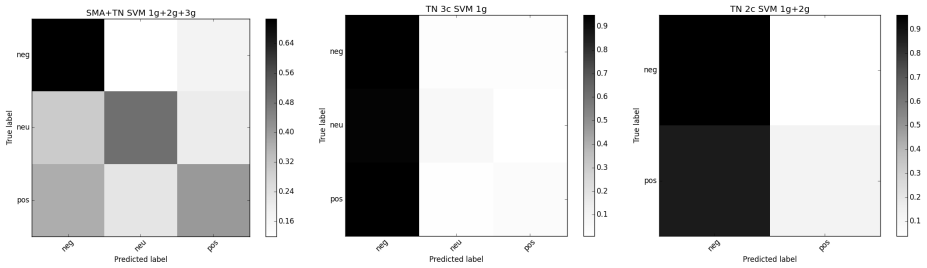
We applied the feature selection and trained the classifiers with features from MSA and TN. Then, we trained the classifiers only on features that are from TN tweets for three classes (positive, negative neutral) and for two classes (positive and negative). The classifiers were run with 10 fold cross validation. The results for the classifiers trained on MSA and TN are presented in Table 6. The best results are obtained with SVM trained with the set of one gram, two grams and three grams as features.

In Table 6 we put the results of the classifiers trained on TN. As Arabic Dialects and specially TN contains words from MSA, we wanted to investigate it has an effect on the accuracy. Looking at the Table 6 we observe an increase in the accuracy of all the classifiers when they are trained only with TN features.

Table 6. The results of the classifiers trained with features from TN. The abreviation g stands for -gram (language model) and c stands for classes.

MSA+TN/ 5514 tweets		**Av. Precision**	**Av. Recall**	**Av. F1**	**Accuracy**
NB	1g	0.50	0.50	0.50	49.34%
SVM	**1g+2g+3g**	**0.54**	**0.53**	**0.53**	**53.55%**
NN	1g+2g	0.50	0.50	0.49	50.11%
RF	1g+2g	0.51	0.51	0.51	51.81%
J48	1g+2g	0.49	0.49	0.47	49.58%
PART	1g	0.50	0.50	0.48	50.08%
TN/ 1754 tweets		**Av. Precision**	**Av. Recall**	**Av. F1**	**Accuracy**
NB 3c	1g	0.49	0.53	0.45	53.42%
NB 2c	1g	0.65	0.69	0.64	69.61%
SVM 3c	**1g**	**0.47**	**0.53**	**0.41**	**53.47%**
SVM 2c	**1g+2g**	**0.67**	**0.71**	**0.63**	**71.09%**
NN 3c	1g	0.47	0.53	0.42	53.24%
NN 2c	1g+2g	0.68	0.70	0.61	70.72%
RF 3c	1g	0.47	0.53	0.43	53.42%
RF 2c	1g+2g+3g	0.67	0.70	0.63	70.87%
J48 3c	1g	0.47	0.53	0.40	53.64%
J48 2c	1g	0.66	0.70	0.62	70.42%
PART 3c	1g+2g	0.40	0.54	0.41	54.04%
PART 2c	1g	0.67	0.71	0.63	71.01%

In Table 6 we show also the results of the classifiers trained on 2 classes (positive and negative). When we reduce the number of classes from 3c (three classes) to 2c (two classes) we increase the accuracy. The best results are obtained withe SVM for the 3c with 1 gram features and for the 2c with 1 gram and 2 gram as features.

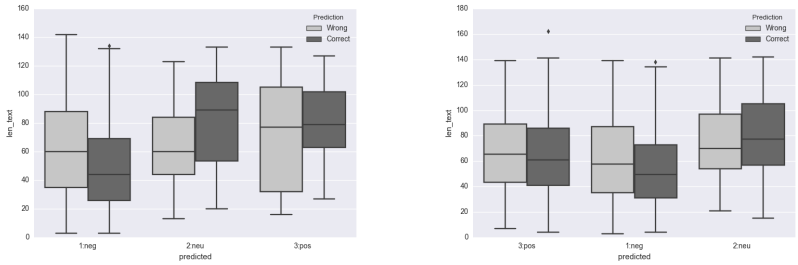


(a) (b) (c)

Fig. 2. Confusion matrix for the best classifiers.

Best classifiers SVM proved to be the best classifier in the three experiments settings. For the best classifiers we plotted the correspondent confusion matrices to see the proportion of miss classified examples and compare it with the given accuracy. Illustrated in Figure 4.2. The first confusion matrix, (a) in Figure 2, illustrates the predicted examples from the SVM trained on MSA and TN. In this confusion matrix we have an important proportion of true negatives and relatively good proportion of true neutrals. The positives are almost equally divided between the true positives and the false positives.

The confusions matrices (b) and (c) are from the SVM trained on TN. (b) is the confusion matrix with 3 classes (positive, negative, neutral). (c) is the confusion matrix with 2 classes (positive, negative). For the two matrices we can see clearly that there is a big proportion of false positives for the positive class and true negatives for the negative class. This means that there is a tendency towards negative class and also a better classification of the negative class, which could be explained by the higher number of negative tweets available for training.



(a) (b)

Fig. 3. Relation between the length of the tweet and the results of the prediction

The length of the tweets and the classification results In this part, we want to investigate the relation between the distribution of the length of the tweets and the predictions made. Figure 4.2 illustrates in different boxes plots, the correct predictions in dark gray and the wrong predictions in light gray. Plot (a) in Figure 3 represents the prediction from the SVM trained on TN and the three classes. Plot (b) in Figure 3 represents the prediction from the SVM classifier trained on SMA and TN.

In Plot (a) we notice that 50% of the correct predictions for the positive class were made on tweets with a length between 60 and 10 characters the wrong predictions include tweets of length less than 80 characters. This differences in predictions open to us a way for different improvement in training the models.

In Plot (b) we notice that the box plots of the correct predictions and the wrong predictions are almost similar. The median length of tweets for the positive

class is around 65 characters. For the tweets trained on MSA and TN we can say that the predictions are barely better than guessing.

5 Conclusion

In this paper, we presented our dataset collected from Twitter in the Tunisian elections context. We manually annotated the tweets in three classes and separated the Modern Standard Arabic and The Tunisian Dialect. We compared the performance of six different classifiers and presented our method for feature selection. We have proposed a benchmark on our dataset that could serve as a base for future works. The observations that we made in this study lead us to some ideas to improve the classification. We intend to extend the dataset and apply further preprocessing methods on Tunisian Dialect to explore if it has any effect on the accuracy of the used classifiers.

References

1. Garrett, R.K., Danziger, J.N.: The internet electorate. Commun. ACM **54** (2011) 117–123
2. Liu, B., Zhang, L.: A Survey of Opinion Mining and Sentiment Analysis. In Aggarwal, C.C., Zhai, C., eds.: Mining Text Data. Springer US (2012) 415–463
3. Adedoyin-Olowe, M., Gaber, M.M., Stahl, F.: A survey of data mining techniques for social media analysis. arXiv preprint arXiv:1312.4617 (2013)
4. Aly, M.A., Atiya, A.F.: LABR: A Large Scale Arabic Book Reviews Dataset. In: ACL (2). (2013) 494–498
5. ElSahar, H., El-Beltagy, S.R.: Building Large Arabic Multi-domain Resources for Sentiment Analysis. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Volume 9042. Springer International Publishing, Cham (2015) 23–34
6. Nabil, M., Aly, M., Atiya, A.F.: ASTD: Arabic Sentiment Tweets Dataset. (In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing) 2515–2519
7. Refaee, E., Rieser, V.: An arabic twitter corpus for subjectivity and sentiment analysis. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14), Reykjavik, Iceland, may. European Language Resources Association (ELRA). (2014)
8. Abdul-Mageed, M., Diab, M., Kbler, S.: SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language **28** (2014) 20–37
9. Abdul-Mageed, M., Diab, M.: Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. Proceedings of the Language Resources and Evaluation Conference (LREC) (2014)
10. Conover, M.D., Gonalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, IEEE (2011) 192–199
11. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Fourth International AAAI Conference on Weblogs and Social Media. (2010)

12. Aiello, L., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., Jaimes, A.: Sensing trending topics in twitter. IEEE Transactions on Multimedia **15** (2013) 1268–1282
13. Mohammed J. Bawaneh, M.S.A.: Arabic Text Classification using K-NN and Naive Bayes. Journal of Computer Science **4** (2008)
14. Abdul-Mageed, M., Diab, M.T., Korayem, M.: Subjectivity and sentiment analysis of modern standard arabic. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Association for Computational Linguistics (2011) 587–591
15. Shaalan, K.: A survey of Arabic named entity recognition and classification. Computational Linguistics **40** (2014) 469–510
16. Salameh, M., Mohammad, S.M., Kiritchenko, S.: Sentiment after translation: A case-study on arabic social media posts. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (2015) 767–777
17. Mohammad, S.M., Salameh, M., Kiritchenko, S.: How Translation Alters Sentiment. Journal of Artificial Intelligence Research **54** (2015) 1–20
18. Monroe, W., Green, S., Manning, C.D.: Word segmentation of informal Arabic with domain adaptation. ACL, Short Papers (2014)
19. Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A.E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.: Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014)
20. Malmasi, S., Refaee, E., Dras, M.: Arabic Dialect Identification using a Parallel Multidialectal Corpus. In: Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015), Bali, Indonesia. (2015)
21. Salzberg, S.L.: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning **16** (1994) 235–240
22. Breiman, L.: Random Forests. Machine Learning **45** (2001) 5–32