Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments

[†]Salima Mdhaffar[1,2], [†]Fethi Bougares[1], [†]Yannick Estève[1] and [‡]Lamia Hadrich-Belguith[2]
[1]LIUM Lab, University of Le Mans, France
[2]ANLP Research Group, MIRACL Lab, University of Sfax, Tunisia
[†]firstname.lastname@univ-lemans.fr
[‡]firstname.lastname@fsegs.rnu.tn

Abstract

Dialectal Arabic (DA) is significantly different from the Arabic language taught in schools and used in written communication and formal speech (broadcast news, religion, politics, etc.). There are many existing researches in the field of Arabic language Sentiment Analysis (SA); however, they are generally restricted to Modern Standard Arabic (MSA) or some dialects of economic or political interest. In this paper we focus on SA of the Tunisian dialect. We use Machine Learning techniques to determine the polarity of comments written in Tunisian dialect. First, we evaluate the SA systems performances with models trained using freely available MSA and Multi-dialectal data sets. We then collect and annotate a Tunisian dialect corpus of 17.000 comments from Facebook. This corpus shows a significant improvement compared to the best model trained on other Arabic dialects or MSA data. We believe that this first freely available[1][2] corpus will be valuable to researchers working in the field of Tunisian Sentiment Analysis and similar areas.

1 Introduction

Sentiment Analysis (SA) involves building systems that recognize the human opinion from a text unit. SA and its applications have spread to many languages and almost every possible domain such as politics, marketing and commerce. With regard to the Arabic language, it is worth noting that the most Arabic social media texts are written in Arabic dialects and sometimes mixed with foreign languages (French or English for example). Therefore dialectal Arabic is abundantly present in social media and micro blogging channels. In previous works, several SA systems were developed for MSA and some dialects (mainly Egyptian and middle east region dialects).

In this paper, we present an application of sentiment analysis to the Tunisian dialect. One of the primary problems is the lack of annotated data. To overcome this problem, we start by using and evaluating the performance using available resources from MSA and dialects, then we created and annotated our own data set. We have performed different experiments using several machine learning algorithms such as Multi-Layer Perceptron (MLP), Naive Bayes classifier, and SVM. The main contributions of this article are as follows: (1) we present a survey of the available resources for Arabic language SA (MSA and dialectal). (2) We create a freely available training corpus for Tunisian dialect SA. (3) We evaluate the performance of Tunisian dialect SA system under several configurations.

The remainder of this paper is organized as follows: Section 2 discusses some related works. Section 3 presents the Tunisian dialect features and its challenges. Section 4 details our Tunisian dialect corpus creation and annotation. In section 4 we report our experimental framework and the obtained results. Finally section 5 concludes this paper and gives some outlooks to future work.

2 Related work

The Sentiment Analysis task is becoming increasingly important due to the explosion of the number of social media users. The largest amount of SA research is carried for the English language, resulting in a high quality SA tools. For many other languages, especially the low resourced ones, an enormous amount of research is required to reach the same level of current applications dedicated

[1]This corpus is freely available for research purpose
[2]https://github.com/fbougares/TSAC

Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 55–61,
Valencia, Spain, April 3, 2017.

to English. Recently, there has been a considerable amount of work and effort to collect resources and develop SA systems for the Arabic language. However, the number of freely available Arabic datasets and Arabic lexicons for SA are still limited in number, size, availability and dialects coverage.

It is worth mentioning that the highest proportion of available resources and research publications in Arabic SA are devoted to MSA (Assiri et al., 2015). Regarding Arabic dialects, the Middle Eastern and Egyptian dialects received the lion's share of all research effort and funding. On the other hand, very small amounts of work are devoted to the dialects of Arabian Peninsula, Arab Maghreb and the West Asian Arab countries. Table 1 summarizes the list of all freely available SA corpora for Arabic and dialects that we were able to find. For more details about previous works on SA for MSA and its dialects, we refer the reader to the extensive surveys presented in (Assiri et al., 2015) and in (Biltawi et al., 2016).

From a technical point of view, the are two approaches to address the problem of sentiment classification: (1) machine learning based approaches and (2) lexicon-based approaches.

Machine learning approaches uses annotated data sets to train classifiers. The sentiment classifier is built by extracting discriminative features from annotated data and applying a Machine learning algorithm such as Support Vector Machines (SVM), Naïve Bayes (NB) and Logistic regression etc. Generally, the best performance is achieved by using n-grams feature, but also Part of speech (POS), term frequency (TF) and syntactic information can be used. (Shoukry and Rafea, 2012) examined two machine learning algorithms: SVM and NB. The dataset is collected from the Twitter social network using its API. Classifiers are trained using unigram and bigram features and the results show that SVM outperforms NB.

Another machine learning approach was used in (Rushdi-Saleh et al., 2011b) where they build the opinion corpus for Arabic (OCA) consisting of movie reviews written in Arabic. They also created an English version translated from Arabic and called *EVOCA* (Rushdi-Saleh et al., 2011b). Support Vector Machines (SVMs) and Naive Bayes (NB) classifiers are then used to create SA systems for both languages. The results showed that both classifiers gives better results on the Arabic version. For instance, SVM gives 90% F-measure on OCA compared to 86.9% on *EVOCA*.

(Abdul-Mageed et al., 2012), have presented SAMAR, a sentiment analysis system for Arabic social media, which requires identifying whether the text is objective or subjective before identifying its polarity. The proposed system uses the SVM-light toolkit for classification.

In lexicon-based approaches, opinion word lexicon are usually created. An opinion word lexicon is a list of words with annotated opinion polarities and through these polarities the application determine the polarity of blocks of text. (Bayoudhi et al., 2015) presented a lexicon based approach for MSA. First, a lexicon has been built following a semi automatic approach. Then, the lexicon entries were used to detect opinion words and assign to each one a sentiment class. This approach takes into account the advanced linguistic phenomena such as negation and intensification. The introduced method was evaluated using a large multi-domain annotated sentiment corpus segmented into discourse segments. Another work has been done in (Al-Ayyoub et al., 2015) where authors built a sentiment lexicon of about 120,000 Arabic words and created a SA system on top of it. They reported a 86.89% of classification accuracy.

3 Tunisian dialect and its challenges

The Arabic dialects vary widely in between regions and to a lesser extent from city to city in each region. The Tunisian dialect is a subset of the Arabic dialects of the Western group usually associated with the Arabic of the Maghreb and is commonly known, as the "Darija or Tounsi". It is used in oral communication of the daily life of Tunisians. In addition to the words from Modern Standard Arabic, Tunisian dialect is characterized by the presence of words borrowed from French, Berber, Italian, Turkish and Spanish. This phenomenon is due to many factors and historical events such as the Islamic invasions, French colonization and immigrations.

Nowadays, the Tunisian dialect is more often used in interviews, telephone conversations and public services. Moreover, Tunisian dialect is becoming very present in blogs, forums and online user comments. Therefore, it is important to consider this dialect in the context of Natural Lan-

Corpus	Size	Language	Source	Reference
ASDT	10000 com	MSA/dialects	Twitter	(Nabil et al., 2015)
OCA	500 doc	MSA	Webpages/Films	(Rushdi-Saleh et al., 2011a)
BBN	1200 com	Levant dialect	Social media	(Zbib et al., 2012)
LABR	63000 com	MSA/dialects	goodreads	(Nabil et al., 2014)
ATT	2154 com	MSA/dialects	TripAdvisor	(ElSahar and El-Beltagy, 2015)
HTL	15572 com	MSA/dialects	TripAdvisor	(ElSahar and El-Beltagy, 2015)
MOV	1524 com	MSA/dialects	elcinema	(ElSahar and El-Beltagy, 2015)
PROD	4272 com	MSA/dialects	souq	(ElSahar and El-Beltagy, 2015)
RES	10970 com	MSA/dialects	qaym	(ElSahar and El-Beltagy, 2015)
Twitter DataSet	2000 com	MSA/Jordanian	Twitter	(Abdulla et al., 2013)
Syria Tweets	2000 com	Syrian	Twitter	(Mohammad et al., 2015)
MASC	8861 com	dialects	Jeeran/qaym/ Twitter/Facebook/ Google Play	(Al-Moslmi et al., 2017)

Table 1: Publically available Arabic SA datasets. Sizes are presented by the number of documents (doc) and commentaries (com).

guage Processing (NLP). The development of SA system for Tunisian dialect faces many challenges due to: (1) the very limited number of previous research conducted in this dialect, (2) the lack of freely available resources for SA in this dialect, (3) and the absence of standard orthographies (Maamouri et al., 2014) (Zribi et al., 2014) and tools dedicated to this dialect.

Indeed, textual content of social networks is characterized by an intense orthographic heterogeneity which made its processing a serious challenge for NLP tools. This heterogeneity is augmented by the lack of normalization of dialectal writing system. Moreover, social networks communication is very impacted by the personal experience of each user. For instance, Tunisian users usually uses code-switching with English or French which depends of their second language.

Table 2 presents an example to highlight the orthographic heterogeneity issue in Tunisian dialect. The example presents the Tunisian dialect translation of the English expression **"how beautiful she is! "**. The translation is a single word which could be written using several spelling variants in Latin or Arabic script in the context of social networks.

Arabic script	Latin script
مَحْلاهَا	Mahleha
مَاحْلاهَا	Ma7lahe
مَاحْلاهَا	Ma7leha
مَا أَحْلاهَا	Ma7laha

Table 2: Example of Tunisian dialect spelling variants of an English expression.

4 Data set collection and annotation

Being aware of the challenges related to the tunisian dialect, we decided to create the first publicly available SA data set for this dialect. This data set is collected from Facebook users comments. Tunisian are among the most active Facebook Users in the Arab Region[3]. In fact, Tunisia is the 8th Arabic country in terms of penetration rates of Tunisian Facebook users, and almost tied as 2nd in the region alongside the UAE (United Arab Emirates) on the percentage of most active users out of total users (Salem, 2017).

This corpus is collected from comments written on official pages of Tunisian radios and TV channels namely *Mosaique FM*, *JawhraFM*, *Shemes FM*, *HiwarElttounsi TV* and *Nessma TV* during a period spanning January 2015 until June 2016.

The collected corpus, called **TSAC** (**T**unisian **S**entiment **A**nalysis **C**orpus), contains 17k user comments manually annotated to positive and negative polarities. Table 4 shows the basic statistics. In particular, we give the number of words, the number of unique words and the average length of

[3] http://www.arabsocialmediareport.com/home/index.aspx

comments per polarity. We provide also the number of Arabic words and mixed comments.

	Positive	**Negative**
# Total Words	63874	49322
# Unique Words	24508	17621
AVG sentence length	7.22	6.00
# Arabic Words	13896	8048
# Mixed comments	98	48
# Comments	8215	8845

Table 3: Statistics of the **TSAC** corpus.

The collected corpus is characterized by the use of informal and non-standard vocabulary such as repeated letters and non-standard abbreviations, the presence of onomatopoeia (e.g. pff, hhh, etc) and non linguistic content such as emoticons. Furthermore, the data set contains comments written in Arabic scripts, Latin scripts known as Arabizi (Darwish, 2014) and even a mixture of both. **TSAC** is a multi-domain corpus consisting of the text covering a maximum vocabulary from education, social and politics domain.

Given the nature of the raw collected data we did some cleaning before the annotation step. We manually : (1) removed the comments that are fully in other languages (French, English, etc.); (2) deleted the user names; (3) deleted URLs and (4) removed hash character from all Hashtags. Table 4, presents several examples for each polarity. We also added the Buckwalter transliteration and the English translation for the purpose of clarity.

5 Experiments and results

From machine learning perspective, the SA could be represented as text classification problem (binary classification in our case). In this section we present several experiments that we run in order to find out (1) the most desirable machine learning algorithms for our task and (2) the usefulness of training data from MSA and other dialects for the Tunisian dialect SA.

5.1 Training Data and features extraction

Table 5 presents the training and evalaution sets. For each corpus we report the dialect, the number of comments per polarity (positive /negative) and the vocabulary size ($|V|$). We used 3 different training corpus, OCA (**O**pinion **C**orpus for **A**rabic), LABR (**L**arge-scale **A**rabic **B**ook **R**eview) and **TSAC**. The OCA corpus contains 500 movie reviews in MSA, collected from forums and websites. It is divided into 250 positive and 250 negative reviews. In this work, we used a sentence level segmented version of OCA corpus described in (Bayoudhi et al., 2015)[4]. The LABR corpus is freely available[5] and contains over 63k book reviews written in MSA and different Arabic dialects. In our experiments we refer to this corpus as mixed dialect corpus (**D_Mix**). The evaluation corpus is a held-out portion, randomly extracted from the **TSAC** corpus to evaluate and compare different SA systems on Tunisian dialect.

In the literature, different linguistic features are generally extracted and successfully used for the SA task. Given the absence of linguistic tools (Part-of-Speech tagger, morphological analysers, lemmatizers, parsers, etc) for Tunisian dialect, we decided to run different classifiers using automatically learned features.

A fixed-length vector is learned in an unsupervised fashion using *Doc2vec* toolkit (Le and Mikolov, 2014) which has been shown to be useful for SA in English (Le and Mikolov, 2014). In this work, each sentence is considered as a document and represented, using *Doc2vec*, by a vector in a multi-dimensional space.

5.2 Classifiers

In SA literature, the most widely used machine learning methods are Support Vector Machines (SVM) and Naive Bayes (NB). On top of these methods, we investigated MLP classifier. All the experiments were conducted in Python using *Scikit* Learn[6] for classification and gensim[7] for learning vector representation. The input of the final sentiment classifier is the set of features vectors from *Doc2vec* toolkit. The output is the sentiment class $S \in \{Positive, Negative\}$.

5.3 SA experiments and evaluation

To evaluate the performance of SA on the Tunisian dialect validation set, we carried out several experiments using various configuration.

Seven experiments were carried out for each classifier depending on the training dataset: (1) using the Tunisian dialect training set, (2) using the

[4]Please contact *Bayoudhi et al.* to obtain a copy of the OCA sentence level segmented corpus

[5]`http://www.mohamedaly.info/datasets/labr`

[6]http://scikit-learn.org/

[7]https://radimrehurek.com/gensim/

Label	Script	Example and Buckwalter transliteration	English translation
Negative	Arabic	mlA hmjyp / ملا همجية	What Savagery
Positive	Arabic	mslsl rwEp / مسلسل روعة	Wonderful series
Negative	Latin	Bsaraha Eni mati3jibnich	Really, I do not like
Positive	Latin	A7sen Moumethel ye3jebni barcha	The best actor, I like it very much
Negative	Mixed	ma8ir ta3li9... فضايح / fDAyH	Scandal...No comment
Positive	Mixed	Bravo صوت رائع / Swp rA}E	Well done great sound

Table 4: **TSAC** annotation examples. Arabic words are given with their Buckwalter transliteration.

		Train set			Evaluation set		
Corpus	Dialect	Positive	Negative	$\mid V \mid$	Positive	Negative	$\mid V \mid$
OCA	MSA	4931	4931	32565	n/a	n/a	n/a
LABR	D_Mix	4880	4880	94789	n/a	n/a	n/a
TSAC	TUN	7145	6515	28480	1700	1700	10791

Table 5: Training corpus. All trained systems are evaluated using the **TSAC** evaluated set.

Classifier	Training set	Positive		Negative		Error rate
		P	R	P	R	
SVM	MSA	0.44	0.15	0.49	0.80	0.52
	D_Mix	0.50	0.84	0.52	0.17	0.49
	TUN	0.77	0.77	0.77	0.76	**0.23**
	MSA_D_Mix	0.51	0.90	0.60	0.15	0.47
	TUN_MSA	0.74	0.83	0.80	0.71	**0.23**
	TUN_D_Mix	0.68	0.76	0.73	0.64	0.30
	ALL	0.71	0.81	0.78	0.66	0.26
BNB	MSA	0.43	0.28	0.46	0.62	0.55
	D_Mix	0.51	0.94	0.58	0.09	0.49
	TUN	0.56	0.70	0.60	0.46	**0.42**
	MSA_D_Mix	0.51	0.98	0.67	0.05	0.49
	TUN_MSA	0.55	0.77	0.62	0.37	0.43
	TUN_D_Mix	0.54	0.76	0.60	0.36	0.44
	ALL	0.54	0.82	0.62	0.30	0.44
MLP	MSA	0.52	0.40	0.51	0.64	0.48
	D_Mix	0.51	0.75	0.53	0.28	0.49
	TUN	0.78	0.78	0.78	0.78	**0.22**
	MSA_D_Mix	0.53	0.49	0.52	0.56	0.47
	TUN_MSA	0.76	0.78	0.77	0.76	0.23
	TUN_D_Mix	0.75	0.77	0.76	0.75	0.24
	ALL	0.74	0.77	0.76	0.73	0.25

Table 6: Results of Tunisian SA experiments using various classifiers with different training sets.

MSA training set, (3) using the mixed MSA and Arabic dialects training set and (4 to 7) using dif-

ferent combination of these datasets.

The performance of our different SA experiments are evaluated on the Tunisian dialect evaluation set and results are reported using precision and recall measures. Precision and recall are defined to express respectively the exactness and the sensitivity of the classifiers.

5.4 Results and Discussion

The results of the different classifiers with different experimental setups are presented in Table 6. As expected, the best classification performance of all the classifiers are obtained when the Tunisian dialect SA system is trained using (or including) the Tunisian dialect training set. We obtained an error rate of 0.23 with SVM, 0.22 with MLP and 0.42 with BNB.

As shown in table 6 SVM and MLP obtain similar results for all experimental setups. However, lower results are obtained with BNB classifier. We notice also no improvement when the SA systems are trained with additional training data from LABR and OCA. Overall, poorer results are obtained when SA systems are trained without the TSAC corpus. This is mainly due to :

- The OCA and LABR data sets are limited to one domain (movies and books respectively), while the evaluation set is multi-domain.
- The OCA and LABR data sets are written only in Arabic character, while the evaluation set contains Latin character.
- The lexical differences between Tunisian dialect, MSA and other dialects.For example, the English word *beautiful*, is written in Tunisian: مزْيَانَة /mizoyaAnap, in Egyptian : حلْوَة / Hilowapo and in MSA : جَمِيلَةٌ / jamiylapN)

Table 7 shows several outputs of our SA system with MLP classifier. We present examples for Positive and Negative classes and for both situation : when SA predict the correct polarity and when SA system fails.

6 Conclusions and feature work

In this paper we have presented the first freely available annotated sentiment analysis corpus for the Tunisian dialect. We have experimented and presented several SA experiments with different training configurations. Best results for Tunisian SA are obtained using the Tunisian training corpus. We believe that this corpus will help to boost research on SA of Tunisian dialect and to explore new techniques in this field. As future works we would like to perform a deep analysis of system outputs. We are planning also to work on the TSAC corpus normalization and to extend the corpus to include the neutral class.

References

Muhammad Abdul-Mageed, Sandra Kübler, and Mona Diab. 2012. Samar: A system for subjectivity and sentiment analysis of arabic social media. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 19–28. Association for Computational Linguistics.

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*, pages 1–6. IEEE.

Mahmoud Al-Ayyoub, Safa Bani Essa, and Izzat Alsmadi. 2015. Lexicon-based sentiment analysis of arabic tweets. *International Journal of Social Network Mining*, 2(2):101–114.

Tareq Al-Moslmi, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. 2017. Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of Information Science*, page 0165551516683908.

Adel Assiri, Ahmed Emam, and Hmood Aldossari. 2015. Arabic sentiment analysis: A survey. *International Journal of Advanced Computer Science and Applications*, 6(12).

Amine Bayoudhi, Hatem Ghorbel, Houssem Koubaa, and Lamia Hadrich Belguith. 2015. Sentiment classification at discourse segment level: Experiments on multi-domain arabic corpus. *Journal for Language Technology and Computational Linguistics*, page 1.

Mariam Biltawi, Wael Etaiwi, Sara Tedmori, and Amjad Hudaib andArafat Awajan. 2016. Sentimnt classification techniques for arabic language: A survey.

Kareem Darwish. 2014. Arabizi detection and conversion to arabic. *ANLP 2014*, page 217.

Hady ElSahar and Samhaa R El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34. Springer.

User comment	System output	Reference
ها الطفل قلولو يسكت امان ناقصين احنا مصاطة	POS	NEG
يا معلم راك ماسط !! تقعد مخيب راسك	POS	NEG
Alah la trabahkom la daniya w la a5ra	POS	NEG
Rajell w m3alem tounsi wakahou	POS	POS
أمنة فاخر أية من الجمال ولكي كل التحية من ليبيا	POS	POS
جعفور يامهبلهوم مشاء الله نشيخ عليك حتي ونتي ساكت	NEG	POS
ممتازة ومتمكنة في التمثيل لكن نريدوا الجديد	NEG	POS
Mitrobi bsaraha .	NEG	POS
5iiiit ech nakraha	NEG	NEG
هذي هي حرية التعبير إلي تحكيو علاها؟؟؟	NEG	NEG

Table 7: Output examples of Tunisian SA system. For each example we present the predicted output and the reference.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development. In *LREC*, pages 2348–2354.

Salameh Mohammad, M Mohammad Saif, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-2015)*.

Mahmoud Nabil, Mohamed A Aly, and Amir F Atiya. 2014. Labr: A large scale arabic book reviews dataset. *CoRR, abs/1411.6718*.

Mahmoud Nabil, Mohamed Aly, and Amir F Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

Mohammed Rushdi-Saleh, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and José M Perea-Ortega. 2011a. Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054.

Mohammed Rushdi-Saleh, Maria Teresa Martín-Valdivia, L Alfonso Ureña-López, and José M Perea-Ortega. 2011b. Bilingual experiments with an arabic-english corpus for opinion mining.

Fadi Salem. 2017. The arab social media report 2017: Social media and the internet of things: Towards data-driven policymaking in the arab world. *Dubai: MBR School of Government.*, 7.

Amira Shoukry and Ahmed Rafea. 2012. Sentence-level arabic sentiment analysis. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pages 546–550. IEEE.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59. Association for Computational Linguistics.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Hadrich Belguith, and Nizar Habash. 2014. A conventional orthography for tunisian arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*, pages 2355–2361.