Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 266–273

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates

Comparative Evaluation of Sentiment Analysis Methods Across Arabic Dialects

Ramy Baly[a], Georges El-Khoury[a], Rawan Moukalled[a], Rita Aoun[a], Hazem Hajj[a], Khaled Bashir Shaban[b], Wassim El-Hajj[c]

[a] *Electrical and Computer Engineering Department, American University of Beirut, Beirut, Lebanon*
[b] *Computer Science and Engineering Department, Qatar University, Doha, Qatar*
[c] *Computer Science Department, American University of Beirut, Beirut, Lebanon*

Abstract

Sentiment analysis in Arabic is challenging due to the complex morphology of the language. The task becomes more challenging when considering Twitter data that contain significant amounts of noise such as the use of Arabizi, code-switching and different dialects that varies significantly across the Arab world, the use of non-textual objects to express sentiments, and the frequent occurrence of misspellings and grammatical mistakes. Modeling sentiment in Twitter should become easier when we understand the characteristics of Twitter data and how its usage varies from one Arab region to another. We describe our effort to create the first Multi-Dialect Arabic Sentiment Twitter Dataset (MD-ArSenTD) that is composed of tweets collected from 12 Arab countries, annotated for sentiment and dialect. We use this dataset to analyze tweets collected from Egypt and the United Arab Emirates (UAE), with the aim of discovering distinctive features that may facilitate sentiment analysis. We also perform a comparative evaluation of different sentiment models on Egyptian and UAE tweets. These models are based on feature engineering and deep learning, and have already achieved state-of-the-art accuracies in English sentiment analysis. Results indicate the superior performance of deep learning models, the importance of morphological features in Arabic NLP, and that handling dialectal Arabic leads to different outcomes depending on the country from which the tweets are collected.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

1. Introduction

Sentiment analysis is the automatic extraction of opinions from words, sentences or documents [28, 26]. It has attracted a lot of attention because of the wide range of applications that can benefit from harvesting the public opinion, and the huge amounts of opinionated data that are available online [37]. In particular, Twitter stands as one of the most used social media platforms, with around 500 million tweets being sent out daily, expressing opinions about personal or trending topics [48, 44]. Tweets are written in all languages including Arabic, which ranks 4th as both the most used language on Twitter [49, 1] and the most spoken language worldwide [35], hence becoming a key source

E-mail addresses: rgb15@mail.aub.edu (Ramy Baly)., gbe03@mail.aub.edu (Georges El-Khoury)., rrm32@mail.aub.edu (Rawan Moukalled)., rra47@mail.aub.edu (Rita Aoun)., hh63@aub.edu.lb (Hazem Hajj)., khaled.shaban@qu.edu.qa (Khaled Bashir Shaban)., we07@aub.edu.lb (Wassim El-Hajj).

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.118

of the Internet content. Given these facts, improving the performance of sentiment analysis models in Arabic tweets is a timely and intriguing problem.

Sentiment analysis generally involves natural language processing (NLP) and machine learning to model text semantics. These tasks are challenging when applied to Arabic tweets for several reasons. First, users tend to tweet using unstandardized Arabic dialects that vary significantly across the Arab world, or using Arabizi; writing Arabic words using Latin characters. They tend to commit misspellings to abide by the allowed space of 140 characters per tweet, and also use special tokens such as hashtags, mentions and URLs, which may express sentiments implicitly. Second, due to the cultural diversity in the Arab world, sentiment models trained on tweets from one region may not applicable to another region. For instance, the phrase *SbHAn Allh* 'God is perfect' is usually used in Levant countries to express positive sentiment, while it is used in Gulf countries to praise God with no sentiment implication. Finally, efforts to create sentiment twitter datasets were limited to specific dialects, mainly Egyptian [38], Gulf [11] and Jordanian [23]. Hence, more corpora are needed to support comparative evaluations of sentiment models across the different Arabic dialects.

In this paper, we describe our efforts to create the Multi-Dialect Arabic Sentiment Twitter Dataset (MD-ArSenTD), which is composed of tweets retrieved from 12 Arab countries. Tweets are assigned sentiment labels according to a 5-point scale to incorporate information of both polarity and intensity. While previous corpora assumed that tweets collected from a specific country are written using that country's dialect [38, 23], MD-ArSenTD contains country and region-level dialect annotation, since we observed that tweets can be written using either Modern Standard Arabic (MSA), the country's dialect, the user's dialect, or a foreign language. This dataset will help the Arabic NLP research community understand the specificities of Arabic tweets by providing insights into Twitter's topics, dialects and writing styles in the different Arab countries. We focus on characterizing tweets from Egypt (from the Nile Basin) and the United Arab Emirates (from the Arabian Gulf). We highlight the differences in the dialects, the discussed topics and the expressed sentiments between tweets of both countries. We also present a comparative evaluation of advanced sentiment models that belong to two classes of machine learning. The first is based on training Support Vector Machines (SVM) using an engineered feature set tailored for Twitter data [22, 14], and the second is based on training deep learning models, namely the Long Short-term Memory networks (LSTM), using generic and dialect-specific embeddings. Results indicate the superiority of deep learning, the importance of using morphological features, and that accounting for the local dialect leads to different outcomes depending on the country from which the tweets are collected. To our knowledge, this is the first attempt to perform sentiment analysis in UAE tweets, and also the first attempt to perform sentiment analysis in Egyptian tweets on a 5-point scale.

The rest of the paper is organized as follows. Section 2 describes previous work on Twitter sentiment analysis. Section 3 describes the system to create the MD-ArSenTD, and presents a characterization study of tweets from Egypt and UAE. Section 4 describes the models used in the comparative evaluation and presents the results along with results analysis. Concluding remarks are provided in Section 5.

2. Related Work

Sentiment analysis in Arabic is usually achieved by training classifiers using different sets of 'engineered' features. The most common features are the word *n*-grams; simple but semantically shallow features that are used to train SVM [42, 10, 45], Naïve Bayes (NB) [32, 24] and ensemble classifiers [34]. Word *n*-grams were also used with syntactic features (root and part-of-speech (POS) tag *n*-grams) and stylistic features (letter and digit *n*-grams, word and document lengths and vocabulary richness), achieving good accuracies after reduction via the Entropy-Weighted Genetic Algorithm (EWGA) [2]. Sentiment lexica provided an additional source of deeper semantic features, which helped boosting the accuracy. Several lexica were developed for MSA (such as ArabSenti [5], ArSenL [12] and SLSA [25]) and for dialectal Arabic (such as SANA [4] and AraSenTi [8]).

Efforts were made to perform sentiment analysis in Arabic Twitter and to create Twitter corpora [9]. A framework was developed to handle Jordanian tweets by training different classifiers using features that capture the semantics of MSA, Jordanian dialect, Arabizi and emoticons [23]. An emoticon-based distant supervision approach improved sentiment classification compared to fully supervised models [39]. A subjectivity and sentiment analysis system for Arabic social media (SAMAR) exploited Arabic morphology using features such as stems, lemmas and POS tag *n*-grams. It also used features such as the presence of polar adjectives, dialect, user ID and gender [3]. Machine