SenZi: A Sentiment Analysis Lexicon for the Latinised Arabic (Arabizi)

Taha Tobaili[1], **Miriam Fernandez**[1], **Harith Alani**[1],
Sanaa Sharafeddine[2], **Hazem Hajj**[3], **and Goran Glavaš**[4]

[1]Knowledge Media Institute, The Open University
[2]Department of Computer Science and Mathematics, Lebanese American University
[3]Department of Electrical and Computer Engineering, American University of Beirut
[4]Data and Web Science Group, Universität Mannheim
{taha.tobaili, miriam.fernandez, h.alani}@open.ac.uk,
sanaa.sharafeddine@lau.edu.lb, hh63@aub.edu.lb, goran@informatik.uni-mannheim.de

Abstract

Arabizi is an informal written form of dialectal Arabic transcribed in Latin alphanumeric characters. It has a proven popularity on chat platforms and social media, yet it suffers from a severe lack of natural language processing (NLP) resources. As such, texts written in Arabizi are often disregarded in sentiment analysis tasks for Arabic. In this paper we describe the creation of a sentiment lexicon for Arabizi that was enriched with word embeddings. The result is a new Arabizi lexicon consisting of 11.3K positive and 13.3K negative words. We evaluated this lexicon by classifying the sentiment of Arabizi tweets achieving an F1-score of 0.72. We provide a detailed error analysis to present the challenges that impact the sentiment analysis of Arabizi.

1 Introduction

Arabizi, a portmanteau of *Arabic* and *Englizi* (English), is a written form of dialectal Arabic (DA) often used by Arabic speakers for informal communication in messaging applications and on social media enabling them to type Arabic words using Latin letters (Yaghan, 2008). Arabizi lacks a consistent orthography and reflects the various dialects of Arabic, which differ from one Arab region to another and from the formal Modern Standard Arabic (MSA) phonetically, morphologically, and syntactically.

Studies show that Arabizi has reached 12% of the Latin script tweets in Lebanon and 25% of the Latin script tweets in Egypt (Tobaili, 2016). It is a common way of communication among the youth (Keong et al., 2015; Muhammed et al., 2011; Allehaiby, 2013) and has been actively used during relevant events in the Arab world such as the Arab spring (Basis-Technology, 2012). Despite the growth of this newly born written language, Arabic sentiment analysis approaches often disregard Arabizi text due to the challenges it poses and the scarcity of NLP resources to process it (Bies et al., 2014).

In this work we contribute to the sentiment analysis of Arabizi by creating a new Arabizi sentiment lexicon (SenZi) for the Lebanese dialect. We annotated a 3.4K Arabizi Twitter dataset to evaluate the lexicon and to train an Arabizi language identifier. We used this identifier to create an Arabizi corpus of 1M public Facebook comments. We widened the coverage of SenZi by enriching it with inflectional and orthographic forms for each sentiment word using word embeddings on the corpus reaching 11.3K positive and 13.3K negative words. All resources and detailed description are made public and freely accessible on the project's webpage[1].

The rest of the paper is structured as follows: Section 2 explains the nature of Arabizi and the challenges it poses. Section 3 reviews the related work. Section 4 presents the annotated datasets and the compiled corpus. Section 5 presents the pipeline for creating SenZi. Section 6 presents the sentiment analysis and results. Section 7 discusses the contributions and limitations of this work. Finally, Section 8 concludes the paper.

2 Background and Challenges

Arabizi privileged its users to transcribe their mother tongue dialect in Latin script at their comfort, free of grammar and orthographic rules. This section dissects the formation of Arabizi, the linguistic issues associated with it, and the challenges it poses onto NLP.

2.1 Background

Arabizi naturally inherits the rich morphology of Arabic but introduces an inconsistent orthography

[1] https://project-rbz.kmi.open.ac.uk

Proceedings of Recent Advances in Natural Language Processing, pages 1203–1211,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_138

and codeswitching.

Morphology: Arabic is an inflectional language where a given word may have a wide range of inflectional forms to express gender, tense, case, number, or perspective. Each of these inflectional forms could be written with different pronoun affixes. An Arabic lemma is inflected by the attachment of clitics, prefixes, and suffixes, the insertion of infixes, or the deletion or replacement of some letters, resulting in a deep morphological shift. For example, the following dialectal words are few inflectional forms of the word زكي *zake* / *smart*: *azkiya, zakeya* /*smart-people* (regular and irregular plural forms), *tetzeka, tetzeke, tetzeko* / *you-are-outsmarting* (masculine, feminine, and plural), *azka* / *smarter-than*, and *ma azkek, azkekon, azkeke* / *how-smart-you-are!* (masculine, feminine, and plural).

Orthography: Arabic is rich in guttural phonemes. It contains two voiceless fricatives خ ح, two voiced fricatives ع غ, a voiced plosive ق, and a glottal stop ء. It also contains distinct consonants with similar phonemes known as soft and emphasised or heavy consonants. Arabic contains five pairs of light and heavy consonants: *q*: ك ق, *t*: ط ت, *d*: د ض, *s*: ص س, and *th*: ذ ظ. This is even exacerbated in Levant Arabic where the ق *q* is pronounced as a glottal stop ء, both ظ and ذ *th* (as in *them*) as ز *z*, and the ث *th* (as in *thrill*) as س *s*. Additionally, there are short and long vowel letters. Short vowels are the diacritics, marks above or below the letters as in كَتَبَ, but they are not scripted in most social texts as in كتب, because a native reader would comprehend the text without the diacritics. These factors had lead to an inconsistent orthography in the transcription of Arabic in Latin script. Moreover, users map the Arabic phonemes with Latin alpha numeral in accordance with their dialect, some transcription standards of the region, and their individual choice of letters. For example:

1. Dialect: The guttural ق *q* is expressed as a guttural *g* in Gulf Arabic but a glottal stop ء in some Levantine Arabic dialects, therefore it is mapped with the number 2 in Levant dialect Arabizi e.g., قلبي *qalbi* / *my heart* in MSA, *galbi* in Gulf Arabizi, and *2albi* or *albi* in Levant Arabizi.
2. Transcription Standards: Some transcriptions became normalised among Arab regions, such as mapping the guttural consonants غ and خ with the numerals 8 and 5 in some countries like Egypt and Jordan (Aboelezz, 2009; Allehaiby, 2013; BIANCHI, 2012) while *gh* and *kh* are more common in Lebanon (Sullivan, 2017).
3. Choice of Letters: It is up to every user's personal choice whether to transcribe some, all, or none of the vowel phonemes either because the short vowels are diacritics or the text is informal and readable without the vowel letters. The following words are few orthographic forms of the word حبيبي *habibi* / *darling* or *my-love*: *7abibi, 7bb, 7bbi, 7abebe, 7bibi, hbb, or habb*.

Codeswitching: Arabizi users constantly switch between Arabizi and Latin script languages, mainly English and French. For example, *Hi! kifak, cava?*, a common trilingual greeting from Lebanon. Codeswitching may occur within individual sentences or within conversations, posing a challenge for data collection and analysis.

2.2 Challenges

Lexical Sparsity: As mentioned earlier (in Section 2.1), Arabizi words can derive a large range of inflectional forms and each form can be transcribed in several orthographic variants. This leads to a high degree of lexical sparsity. Therefore, sentiment lexicons with one or few forms for each sentiment word are insufficient to capture the high number of inflections and variants that could be derived from each Arabizi word.

Word Ambiguity: Apart from words that are naturally polysemous, transcribing Arabic phonemes that have no equivalent in English Latin script may lead to ambiguity. Ambiguous words are generated by:

1. Transcribing a short Arabic vowel phoneme, a diacritic, as a vowel letter in Latin script. For example, transcribing the word ضيعة (short vowel *a*, a diacritic originally) / *village* as *day3a* becomes ambiguous with ضايعة (long vowel *a*) / *lost* or *confused*.
2. Transcribing one Latin script letter for two distinct Arabic letters. This is common for the soft and heavy consonants. For example, transcribing درب (soft *d)* / *route* as *dareb* be-

comes ambiguous with ضرب (heavy *d*) / *hit*.

Transliteration: Transliteration in this context is the automatic conversion of Arabizi into Arabic script. With the heterogeneity of Arabic dialects and the inconsistency of orthography, transliteration can not be achieved in a straightforward Latin to Arabic mapping. The most accurate transliterators are online tools that generate a list of possible transliterations for every input word such as Yamli[2] and Google Input Tools[3]. These tools are designed to help Arabic users output MSA text by typing in Latin script Arabic word by word. Transliterating whole Arabizi texts to Arabic produces orthographic errors. Google Translate for example, detects Arabizi, converts it to Arabic, and translates it to the target language. It translated the tweet *da5l jamelik w hadamtik / Oh-my (expression), your-beauty and your-humour (feminine)* to *inside your camel and demolished* due to the dialect (Lebanese), choice of letters, and word ambiguity.

3 Literature Review

Recent efforts in creating lexical resources for Arabic focus mainly on MSA (Badaro et al., 2014; Eskander and Rambow, 2015; Al-Twairesh et al., 2016) and DA (Abdul-Mageed and Diab, 2014). Several works that analysed sentiment from Arabic social data filtered out Arabizi completely from their datasets (Al-Kabi et al., 2013, 2014; Duwairi and Qarqaz, 2014), missing the sentiment from a considerable portion of the population in general and the youth in specific.

To the best of our knowledge, (Duwairi et al., 2016; Mataoui et al., 2016; GUELLIL et al., 2018) are the only works that looked into sentiment analysis for Arabizi. All three papers proposed to transliterate Arabizi to Arabic. (Duwairi et al., 2016) transliterated Jordanian dialect Arabizi to Arabic using their own transliterator without evaluating the transliterations. (Mataoui et al., 2016) transliterated Algerian dialect Arabizi as part of their sentiment analysis for Algerian Arabic pipeline. They used Google Translate without evaluating the transliterations as well. (GUELLIL et al., 2018) created a rule-based transliterator that generates several transliterations per word, then used a language model based on a large corpus to select the best transliteration. Thus, minimising the error of transliteration but maximising the complexity of the task.

The following papers propose sophisticated works for transliterating Egyptian (Darwish, 2014; Al-Badrashiny et al., 2014; Eskander et al., 2014) and Algerian dialect Arabizi (Guellil et al., 2017). The limitations of these works include transliterating datasets manually, hand-crafting rules to preprocess and normalise Arabizi, and mapping Arabizi with Arabic heuristically.

In this work we aim to advance the state of the art in Arabic sentiment analysis by analysing Arabizi directly, without the need to filter, preprocess, or transliterate it.

4 Data Collection and Annotation

This section describes the collection and annotation of Twitter datasets to evaluate SenZi and train an Arabizi identifier. It then describes the creation of a Facebook corpus that was used (in Section 5.2) to enrich SenZi with inflectional and orthographic forms.

4.1 Annotated Twitter Datasets

Collection: We used the Twitter stream API to collect live tweets that have geographic coordinates lying within the region of Lebanon. We collected 177K tweets intermittently over the period of one year, between 2016 and 2017. We filtered out the Arabic tweets that were identified by Twitter as Arabic: ar (80K). The remaining dataset contains 97K tweets in Latin script such as, Arabizi, English and French or codeswitched tweets among these languages. Twitter misidentified Arabizi and codeswitched tweets as (ht, tr, in, hi, pt, nl, ct, ey) where some stand for (Haitian, Turkish, Hindi, Portuguese, and Dutch). To accurately identify the Arabizi tweets, we resorted for a manual annotation task.

Preprocessing: We removed the URLs, hashtags, mentions, and non-ASC characters and deleted duplicated tweets and tweets that lack an alphabet, obtaining a filtered dataset of 66K tweets.

Annotation: We selected 30K tweets randomly and created a user friendly annotation platform that displays these tweets in different random order for every user. The platform asks each user:

1. Is the tweet written mostly in Arabizi?
   (*yes, no, I don't know*)

[2] `http://www.yamli.com`

[3] `http://www.google.com/inputtools/try`

2. What is the sentiment of the tweet? (*positive, negative, neutral, I don't know*)

We assigned this task to three Lebanese undergraduate students and guided them to:

1. Count the number of Arabizi words in codeswitched tweets to determine if these tweets are mostly written in Arabizi.
2. Consider the sentiment that the tweets infer regardless of present expressions e.g., *haha tabashna bl exam* / *haha we failed the exam* is a negative tweet.
3. Answer *I don't know* for ambiguous tweets.

A screenshot of the developed annotation platform is presented in Figure 1. Further details about the platform and the annotation process can be found on the project's webpage.

Results: Table 1 presents the annotation of the 30K tweets for Question 1: *Is the tweet written mostly in Arabizi?* We had a total of 4.3K *yes*, 27.6K *no*, and 641 *I don't know*. We applied Fleiss' Kappa (Fleiss, 1971) to measure the agreement among the annotators scoring a substantial agreement of 0.74 (Landis and Koch, 1977).

Table 1: Arabizi Annotation of 30K Tweets

Tweets	Arabizi	Not Arabizi	IDK	Kappa
30K	4.3K	27.6K	641	0.74

From the 4.3K Arabizi tweets, there were (3.4K tweets) where at least two answers match for *Arabizi-yes* and (2.2K Tweets) where all three answers match for *Arabizi-yes*. We balanced the 2.2K Arabizi with a 2.2K non-Arabizi tweets to create an Arabizi identification (AI) dataset.

Table 2 presents the annotation of the (3.4K tweets) for Question 2: *What is the sentiment of the tweet?* We had a total of 1.2K *positive*, 1.4K *negative*, 2.1K *neutral*, and 172 *I don't know* scoring a fair agreement of 0.33 Fleiss' Kappa.

Table 2: Sentiment Annotation of the (3.4K Tweets)

Tweets	Pos	Neg	Neutral	IDK	Kappa
3.4K	1.2K	1.4K	2.1K	172	0.33

From the 3.4K Tweets, there were (2.9K Tweets) where at least two answers match for the sentiment of the Arabizi tweets. They consist of 801 *positive*, 881 *negative*, 1.2K *neutral*, and 7 *I don't know*. We balanced an 800 positive with 800 negative tweets to create the sentiment analysis (SA) dataset.

As a result, we had two datasets:

1. AI Dataset: 4.4K Tweets (2.2K Arabizi and 2.2K not Arabizi).
2. SA Dataset: 1.6K Arabizi Tweets (800 positive and 800 negative).

We used the AI Dataset to train the Arabizi identifier (in Section 4.2) and the SA Dataset to evaluate SenZi (in Section 6.1).

4.2 Automatic Arabizi Identification

We used the AI Dataset (from Section 4.1) to train a Support Vector Machine (SVM) classifier with the tweets' unigrams as input features. We shuffled the dataset and split it into 10 folds for cross validation. The average of the classification results for all folds are presented in Table 3.

Table 3: Arabizi Identification
4.4K Tweets: 2.2K Arabizi - 2.2K non-Arabizi

Recall	Precision	F1-score	Accuracy
0.93	0.97	0.95	0.95

4.3 Facebook Corpus

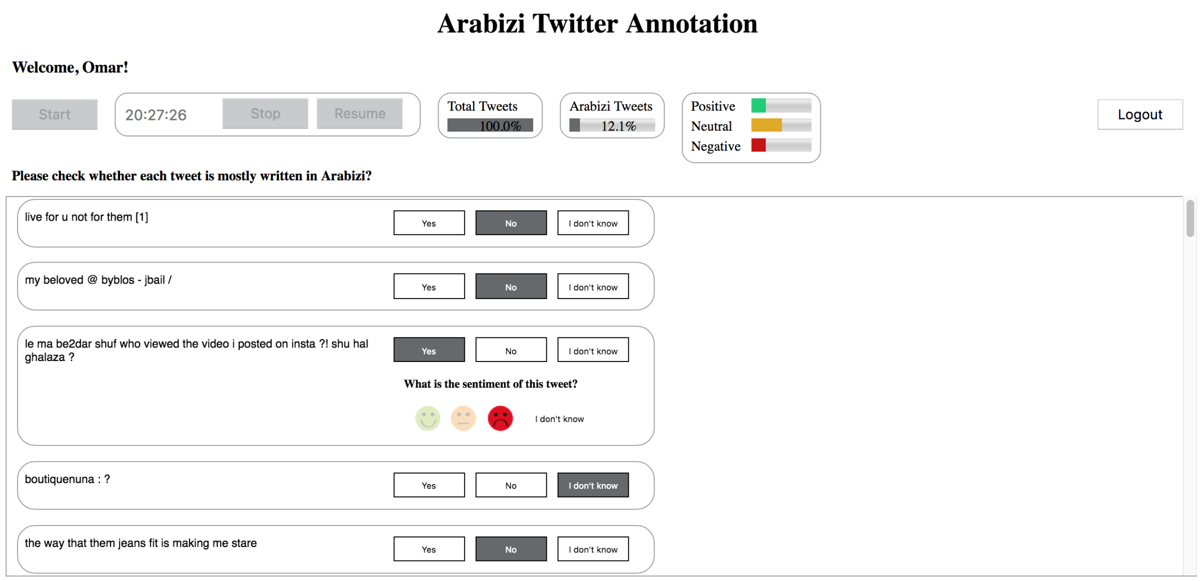
We created an Arabizi corpus of 1M Facebook public comments collected from 49 popular and active Lebanese pages[4]. The pages vary in genre such as, news, comedy, and politics. We used this corpus (in Section 5.2) to create a word embeddings space to discover inflectional and orthographic forms of SenZi's sentiment words.

Collection: We wrote a script that uses Facebook API to iterate over all posts (texts, images, and videos) in a public page and extract all the comments and replies from every post. It collects all Latin script comments and replies in reverse chronological order up to the very first post posted by the page. As we extracted the comments and their replies, we skipped comments that contain Arabic text. We ran the script over the 49 public pages in 2017 resulting in a 2.2M Latin script Facebook comments.

Preprocessing: We removed the URLs, mentions, and media attachments and deleted duplicated comments and comments that lack an alphabet, reducing the comments to 2.1M.

Identification: We used the trained Arabizi identifier (from Section 4.2) to identify the Arabizi text, obtaining a corpus of 1M Arabizi Facebook comments.

[4] The list of pages can be found at the project's webpage.



Arabizi Twitter Annotation

Welcome, Omar!

Total Tweets 100.0%

Arabizi Tweets 12.1%

Positive 🟩
Neutral 🟨
Negative 🟥

Please check whether each tweet is mostly written in Arabizi?

live for u not for them [1] — Yes [✓ No] I don't know

my beloved @ byblos - jbail / — Yes [✓ No] I don't know

le ma be2dar shuf who viewed the video i posted on insta ?! shu hal ghalaza ? — [✓ Yes] No I don't know

What is the sentiment of this tweet? 🙂 😐 [✓ ☹] I don't know

boutiquenuna : ? — Yes No [✓ I don't know]

the way that them jeans fit is making me stare — Yes [✓ No] I don't know

Figure 1: Annotation Platform

5 SenZi

This section presents the pipeline of building SenZi. It was built in two phases:

1. Lexicon Generation: Using existing resources to generate an initial list of Arabizi sentiment words.
2. Lexicon Expansion: Expanding this sentiment word list using the created Facebook corpus (from Section 4.3).

5.1 Lexicon Generation

Resources: We started with two English sentiment lexicons and one Lebanese dialect word list as seeds to build SenZi. We chose two of the most common English sentiment lexicons in the literature: Hiu and Liu[5] (2K positive and 4.8K negative) and the MPQA[6] (2.7K positive and 4.9K negative) (Wilson et al., 2005). LivingArabic is a list of 7.1K Lebanese dialect words compiled by the Living Arabic project[7]. We generated the Lebanese Arabizi sentiment words in five steps:

1. Combine the existing English lexicons.
2. Translate to Arabic.
3. Select the dialectal sentiment words.
4. Combine the resulting Arabic lexicons.
5. Transliterate to Arabizi.

5.1.1 Combination: English Lexicons

We took the union of Hiu Liu and MPQA to have a list of 2.7K positive and 5.1K negative words. We call this list HL-MPQA.

5.1.2 Translation

We used an online dictionary *bab.la*[8] to translate HL-MPQA to MSA. We wrote a script that inputs every word from HL-MPQA into *bab.la* and copies the single-word translations keeping the multi-word expressions for a future work. We generated 4.2K positive and 5.2K negative unique MSA words. We call this list HL-MPQA-Ar.

5.1.3 Selection

HL-MPQA-Ar: Since we aimed to create a Lebanese dialect lexicon, we needed to filter HL-MPQA-Ar from terms that are not common to the Lebanese dialect. We asked a Lebanese graduate student to select the dialectal sentiment words. The student selected 537 positive and 1K negative dialectal terms.

LivingArabic: We assigned an annotation task to three undergraduate Lebanese students to select the sentiment words from LivingArabic. The students selected 531, 672, and 1K sentiment words. We took 732 words (179 positive and 553 negative) where at least two students had agreed on the sentiment.

[5] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[6] https://mpqa.cs.pitt.edu/lexicons/subj_lexicon

[7] http://www.livingarabic.com

[8] https://bab.la

5.1.4 Combination: Arabic Lexicons

We took the union of the selected dialectal words from HL-MPQA-Ar and the selected sentiment words from LivingArabic to have a list of 607 positive and 1.4K negative Arabic words.

5.1.5 Transliteration

We asked the Lebanese graduate student to transliterate the resulting dialectal Arabic words to Arabizi. This marked the first version of SenZi, consisting of 2K Arabizi words (607 positive and 1.4K negative).

5.2 Lexicon Expansion

As mentioned in Section 2, Arabizi inherits the rich morphology of Arabic and introduces an inconsistent orthography causing a high degree of lexical sparsity. We addressed this challenge by expanding SenZi automatically to cover a range of inflectional and orthographic forms for each sentiment word.

We created a word vector space by training a word embeddings fastText skipgram model (Bojanowski et al., 2016) on the 1M Arabizi FB comments corpus (from Section 4.3). We retrieved a vector of nearest word neighbours for every SenZi word. Then, we matched the inflectional and orthographic forms from the retrieved vector with the SenZi word automatically. We learned heuristically that a retrieved Arabizi word is a form of a SenZi word if it contains the same sequence of consonant letters of that SenZi word. We maximised the inflectional and orthographic forms from the retrieved vector by normalising its words lightly to stay consistent with the words of SenZi. We ran this expansion twice, recursively.

5.2.1 First Expansion:

1. Retrieve a vector of 50 nearest word neighbors for each SenZi word.
2. Normalise the orthography of these words:
   - Replace *8, 5* and *ch* with *gh, kh* and *sh*.
   - Remove repeated letters (exaggeration).
3. Take the nearest word neighbors (in their original form, before normalisation) that contain the same consonant letter sequence of the SenZi word.

For example, the consonant letter sequence *tyb* from the word *tayab* / *cute* or *tasty* matched 30 inflectional and orthographic forms such as, *atyab, atyabak, atyabo, 2tyab, tayoub, taybe, tayoubi, taybeee, taybin,* and *tayoubin,*.

We note that applying this technique against all words in the corpus would match many irrelevant words because most Arabic words stem from triliteral words, but in this case we are limited with the nearest word neighbors, i.e., words that are semantically related.

This expanded SenZi to 12.3K words (5.1K positive and 7.2K negative).

5.2.2 Second Expansion:

1. Retrieve a vector of 50 nearest word neighbors for each *new* word.
2. Normalise the orthography of these words.
3. Take the nearest word neighbors that contain the same consonant letter sequence of the *original* SenZi word (not the new word).

For example, the word *atyabak* was retrieved in the first expansion of the word *tayab*. In the second expansion we match the consonant letter sequence *tyb* of the original word *tayab* with the word neighbors of the newly retrieved word *atyabak*. This further expanded *tayab* with two new inflectional forms of *atyabak* in two different orthographies *atyabek, atyabik, atyabkon,* and *atyabkoun* / *cute-singular-feminine* and *cute-plural* in the second person perspective.

We cleaned SenZi by deleting duplicates of words and words that occurred in both the positive and negative lists. This further expanded SenZi to 24.6K words (11.3K positive and 13.3K negative).

6 Evaluation

This section describes the sentiment analysis approach and results and presents a manual error analysis.

6.1 Evaluation Setup and Results

We followed the evaluation method of AraSenti, an Arabic sentiment lexicon proposed by (Al-Twairesh et al., 2016), to evaluate SenZi. we applied a 2-class sentiment classification using a simple lexicon-based approach. We evaluated SenZi before and after the expansion against the SA Dataset (from Section 4.1) which consists of 800 positive and 800 negative tweets.

We created a list of 10 negators and expanded it to 170 words[9] using the same lexicon expansion technique. We classified a sentiment word in its opposite sentiment class if it is preceded by a

[9] List of negators available at the project's webpage.

negator. However, the negator ما *ma* acts as an intensifier in some cases. For example, the *ma* in *ma ajmala! / How beautiful-she/it-is!* precedes an inflection of *jamil / pretty* that begins with a glottal stop أ *2* or *a*, we therefore exempted this negator from negation if followed by a glottal stop.

We matched the positive and negative words in the tweets with SenZi and classified the tweet *positive* if the positive matches were greater than the negative matches, *negative* if the negative matches were greater than the positive matches, and *no sentiment* otherwise. Since this is a 2-class classification and the dataset is balanced, we classifed tweets with *no sentiment* as *positive* or *negative* randomly. The average results are presented in Table 4.

Table 4: SenZi Evaluation
Lexicon-Based Classification

	R	P	F	A
SenZi Original	0.56	0.59	0.57	0.58
SenZi 1st Expansion	0.74	0.64	0.69	0.67
SenZi 2nd Expansion	0.79	0.66	0.72	0.69

Enriching SenZi with inflectional and orthographic forms pushed the F1-score by a solid 0.15 over the baseline. This implies that the word forms and variants play a significant role in sentiment analysis for Arabizi.

6.2 Error Analysis

We provide a detailed error analysis for the lexicon-based classification using the best version of SenZi (2nd expansion) in classifying 800 positive and 800 negative tweets (SA Dataset from Section 4.1). The confusion matrix of this classification is presented in Table 5.

Table 5: Confusion Matrix
Lexicon-Based Classification

Actual	Classified		
	Positive	Negative	No Sentiment
Positive	55%	3%	42%
Negative	13%	39%	48%

Most of the error (45%) lies in not determining a sentiment class for the tweets. As such, we extracted a sample of 100 positive and 100 negative tweets that were wrongly classified for a manual assessment. We checked all the words in the sample whether matched or missed by SenZi. The challenges that impact the performance of the lexicon-based approach and their percentages are presented in Table 6 followed by a small discussion for each challenge.

Sentiment form not in the lexicon: The main factor for not classifying sentiment tweets was due to lexical sparsity, the same challenge that we addressed in this paper. Although we expanded SenZi from 2K to 24.6K words with an increase of 0.23 in recall, it still did not match 38% of inflectional and orthographic forms of SenZi's sentiment words.

Sentiment word is in English: Although the Twitter dataset was annotated as mostly Arabizi, 12% of the unclassified sentiment words are written in English. Sentiment words in English appeared in the positive set slightly more than the negative set with expressions like *my love, miss you, happy birthday, best wishes,* and *good luck* over cursing and swearing in the negative set. Borrowing is also common in Arabizi e.g., *luvik* and *missik* for *love-you-feminine* and *miss-you-feminine* in the second person perspective.

Neutral word classified: The drawback of the automatic expansion of words is a decrease in precision with 14% wrong classification of neutral words in this case.

Multi-word expressions and sarcasm: Many common multi-word expressions that express sentiment or sarcasm lack sentiment words, hence bypass a simple lexicon-based approach. For example, *to2bor albe / burry my heart* expresses love or *ras kbeer / big head* means stubborn.

No sentiment words: 9% of the unclassified tweets lack sentiment words with a higher tendency in the negative class. For example, the translated negative tweet *mom woke me up 30 minutes ago saying common common you have to give your sister a ride, guess who is still waiting?* or the positive *lets listen to keaton henson and eat shawarma*. This is an open problem in the literature of sentiment analysis (Liu, 2012).

Sentiment word not in the lexicon: SenZi did not match 6% of the unclassified sentiment words.

Word Ambiguity: We identified 5% of the wrongly classified words as ambiguous. As mentioned previously (in Section 2.2), word ambiguity is one of the Arabizi NLP challenges generated by the transcription of Arabic in Latin script.

Wrong negation: Classifying negated sentiment words accurately requires more effort than

Table 6: Challenges of Arabizi Sentiment Analysis

	Sentiment form not in the lexicon	Sentiment word is in English	Neutral word classified	Multiword expressions or sarcasm	No sentiment words	Sentiment word not in the lexicon	Word ambiguity	Wrong negation
Positive	37.5%	15%	12.5%	11%	5%	8%	5%	3%
Negative	39%	10%	15.5%	10%	12%	4%	5%	4%

negating sentiment words that are preceded by a negator. A negator may precede or succeed a sentiment word by several words and it may only diminish the sentiment in some cases.

7 Discussion

In this work we focused on an area that has not been explored thoroughly in the literature of sentiment analysis. Arabizi has been proven to be a prominent way of texting on social media among the Arab youth yet there are no public resources to analyse sentiment from this script.

We provided a rigorous explanation of the linguistic challenges for analysing Arabizi text. We created SenZi, the first Lebanese Arabizi sentiment lexicon. We addressed the high degree of lexical sparsity by enriching SenZi with different inflectional and orthographic forms using word embeddings. We achieved an F1-score of 0.72 using a lexicon-based sentiment classification approach.

To the best of our knowledge, there are no other Levant dialect Arabizi datasets or sentiment lexicons to compare our work with. As such, we provided a detailed error analysis to point out the cases that bypassed SenZi.

The annotations carried out to create SenZi and the datasets took place at different times between 2016 and 2018. We tried our best to keep three annotators per task, but in a few cases we had one annotator at hand. However, we tested the annotators with test sets for credibility.

Word embeddings proved to be an excellent technique to expand SenZi, yet 38% of the unmatched sentiment words are forms of SenZi words. Next, we will explore cross-lingual word embeddings with Arabic for further expansion.

Arabizi is a code-switched language, with English appearing the most in Arabizi text from Lebanon. We plan to add English sentiment words to SenZi carefully to handle codeswitching.

Nevertheless, this work is one of the very first attempts to create and evaluate NLP resources for Arabizi sentiment analysis. We created a new sentiment lexicon consisting of 11.3K positive and 13.3K negative words, a sentiment-annotated dataset of 3.4K tweets, and a Facebook corpus of 1M comments. All resources and detailed description are made public and freely accessible on the project's webpage[10].

8 Conclusion

We presented SenZi, the first sentiment analysis lexicon for the Lebanese dialect Arabizi. We built it by translating, annotating, and transliterating other resources to have an initial set of 2K sentiment words. We expanded it to 24.6K sentiment words by importing inflectional and orthographic forms using word embeddings. We evaluated it using a lexicon-based sentiment analysis, achieving an F1-score of 0.72. We finally presented a detailed error analysis to pinpoint its limitations and the challenges that impact the lexicon-based approach for Arabizi sentiment analysis.

Acknowledgements

We thank the annotators and the web developers for their fundamental contributions.

References

Muhammad Abdul-Mageed and Mona T Diab. 2014. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *LREC*, pages 1162–1169.

Mariam Aboelezz. 2009. Latinised arabic and connections to bilingual ability. In *Papers from the Lancaster University Postgraduate Conference in Linguistics and Language Teaching*.

Mohamed Al-Badrashiny, Ramy Eskander, Nizar Habash, and Owen Rambow. 2014. Automatic transliteration of romanized dialectal arabic. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 30–38.

Mohammed Al-Kabi, Amal Gigieh, Izzat Alsmadi, Heider Wahsheh, and Mohamad Haidar. 2013. An opinion analysis tool for colloquial and standard arabic. In *The Fourth International Conference on Information and Communication Systems (ICICS 2013)*, pages 23–25.

[10] https://project-rbz.kmi.open.ac.uk

Mohammed N Al-Kabi, Amal H Gigieh, Izzat M Alsmadi, Heider A Wahsheh, and Mohamad M Haidar. 2014. Opinion mining and analysis for arabic language. *International Journal of Advanced Computer Science and Applications (IJACSA), SAI Publisher*, 5(5).

Nora Al-Twairesh, Hend Al-Khalifa, and Abdulmalik AlSalman. 2016. Arasenti: Large-scale twitter-specific arabic sentiment lexicons. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 697–705.

Wid H Allehaiby. 2013. Arabizi: An analysis of the romanization of the arabic script from a sociolinguistic perspective. *Arab World English Journal*, 4(3):52–62.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. In *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*, pages 165–173.

Basis-Technology. 2012. The burgeoning challenge of deciphering arabic chat.

Robert Michael BIANCHI. 2012. 3arabizi-when local arabic meets global english. *Acta Linguistica Asiatica*, 2(1):89–100.

Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander, and Owen Rambow. 2014. Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 93–103.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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

Rehab M Duwairi, Mosab Alfaqeh, Mohammad Wardat, and Areen Alrabadi. 2016. Sentiment analysis for arabizi text. In *2016 7th International Conference on Information and Communication Systems (ICICS)*, pages 127–132. IEEE.

Rehab M Duwairi and Islam Qarqaz. 2014. Arabic sentiment analysis using supervised classification. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, pages 579–583. IEEE.

Ramy Eskander, Mohamed Al-Badrashiny, Nizar Habash, and Owen Rambow. 2014. Foreign words and the automatic processing of arabic social media text written in roman script. *EMNLP 2014*, page 1.

Ramy Eskander and Owen Rambow. 2015. Slsa: A sentiment lexicon for standard arabic. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2545–2550.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Imane GUELLIL, Ahsan Adeel, Faical AZOUAOU, Ala-eddine Hachani, Amir Hussain, et al. 2018. Arabizi sentiment analysis based on transliteration and automatic corpus annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 335–341.

Imane Guellil, Faiçal Azouaou, Mourad Abbas, and Sadat Fatiha. 2017. Arabizi transliteration of algerian arabic dialect into modern standard arabic. In *Social MT 2017/First workshop on Social Media and User Generated Content Machine Translation*.

Yuen Chee Keong, Othman Rahsid Hameed, and Imad Amer Abdulbaqi. 2015. The use of arabizi in english texting by arab postgraduate students at UKM. *The English Literature Journal*, 2(2):281–288.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

M'hamed Mataoui, Omar Zelmati, and Madiha Boumechache. 2016. A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Res. Comput. Sci*, 110:55–70.

Randa Muhammed, Mona Farrag, Nariman Elshamly, and Nady Abdel-Ghaffar. 2011. Summary of arabizi or romanization: The dilemma of writing arabic texts. In *Jīl Jadīd Conference, University of Texas at Austin*, pages 18–19.

Natalie Sullivan. 2017. *Writing Arabizi: Orthographic Variation in Romanized Lebanese Arabic on Twitter*. Ph.D. thesis.

Taha Tobaili. 2016. Arabizi identification in twitter data. *ACL 2016*, page 51.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Mohammad Ali Yaghan. 2008. "arabizi": A contemporary style of arabic slang. *Design Issues*, 24(2):39–52.