Lexical Induction of Morphological and Orthographic Forms for Low-Resourced Languages

Taha Tobaili
Knowledge Media Institute
The Open University
taha.tobaili@open.ac.uk

Abstract

Languages with large vocabularies pose a challenge for lexical induction especially if they are too low in resources to train sophisticated language models. The situation becomes even more challenging if the language lacks a standard orthography, such that spelling variants cannot be referenced with a standard format for lexical simplification. We propose a simple approach to induce morphological and orthographic forms in low-resourced NLP and prove its potential on a non-standard language as a use case.

1 Introduction

A language is considered to have a high degree of lexical sparsity if the words of the language could have a high number of forms (Baly et al., 2017), whether morphological (inflections) or orthographic (spellings). As a result the number of possible forms for many words becomes large, a considerable challenge in NLP for low-resourced languages. With the lack of linguistic resources, several NLP tasks become difficult to achieve such as named entity recognition (NER), part of speech (POS) tagging, parsing, question answering (QA), machine translation (MT), and sentiment analysis.

Stemming, lemmatisation, and normalisation of text are the most intuitive text preprocessing tasks to address a highly-sparsed language[1], with the goal of simplifying a variety of word forms into one standard form, such that the inflected or differently-spelled forms match easily with their lemmas or roots in a lexicon or corpus. On the other hand, it reduces the number of word forms in a dataset to train a machine learning (ML) algorithm. As plausible as this direction sounds, it is not as straightforward for morphologically-rich languages (MRLs) as it is for languages with simple morphology. In English for example, the words *enjoyed, enjoying, enjoyable, joyful, and unjoyful* can easily map with the lexemes *enjoy* and *joy*. In Arabic however, the inflections are not limited to prefixes and suffixes but also include infixes, proclitics, and enclitics. The affixes consist of objects and pronouns that are tense, number, and gender sensitive as well, that one verb could easily expand to a hundred inflectional forms. Arabic is also rich in tri-literal words because of the short vowel diacritic transcript, as such even a slight inflection could map with a list of irrelevant words, for example: استنكار (*denying*) maps with eleven words of different meanings سكر نكر نار تنكر تنك كر سار تار ستار ستر استنار.

Mapping words with their base forms becomes even more difficult when the language is purely spoken, lacking a standard orthography. Thus, if transcribed, people spell words according to their own perception of how phonemes should be represented in text, resulting in various spellings for each word.

In this work, we choose Arabizi (Yaghan, 2008), an informal heuristic transcription of the spoken dialectal Arabic (DA) in Latin script. A language that poses both issues, rich morphology and inconsistent orthography. Not only can each word be inflected into many forms, but also each form could be written in a variety of ways. Hence, the magnitude of its lexical sparsity quickly becomes very large as it is a multiple of its morphology and orthography.

[1] A language with a high degree of lexical sparsity.

Proceedings of the Third Workshop on Multilingual Surface Realisation (MSR'20), pages 42–49
Barcelona, Spain (Online), December 12, 2020.

Unlike the formal Modern Standard Arabic (MSA), the spoken vernacular DA is quiet esoteric to different regions with different structure, word choice, vocabulary, conjugation, speech tempo, and pronunciation. It is influenced by other languages as well such as French, Turkish, Spanish, or Italian based on the history of the region. We focus on the Lebanese dialect Arabizi as the use case of this study, a member of the larger Levantine dialect family with less resources in the literature than Egyptian, North-African, or Gulf dialects. The Lebanese dialect is also known for its mixing with English and French which is also reflected in its Arabizi transcription (Sullivan, 2017; Tobaili, 2016), another challenge for word classification.

We address the problem of the large lexical sparsity of Arabizi by taking a reversed approach; rather than attempting to lemmatise or simplify words, we retrieve a high number of written forms for given input words from a raw corpus. We propose an approach that combines rule-based mapping with word embeddings to automatically expand Arabizi words. We test this approach on a lexicon composed of 2K sentiment words that resulted in a new sentiment lexicon of 175K words. We evaluate the approach extrinsically by testing the newly expanded sentiment lexicon against a pre-annotated Arabizi dataset.

As such, the contributions of this work are mainly an approach to find the written forms of words in highly-sparsed low-resourced languages and the outcome resource, a new Lebanese dialect Arabizi sentiment lexicon.

2 Related Work

In this work we build upon our previous work of (Tobaili et al., 2019), where we compiled a raw corpus and developed a new sentiment lexicon and datasets for the Lebanese dialect Arabizi. We started by developing a 2K sentiment words lexicon and followed by expanding it using the cosine word similarity feature, nearest neighbours, from an embedding space trained on the raw corpus. This yielded in a lexicon of 24.6K words, named SenZi, an average of 12 forms per word, to cover some inflectional and orthographic variants of the original sentiment words.

However, the degree of the lexical sparsity in Arabizi is higher than 12 variants per word as can be seen in Section 4. We therefore, propose a new approach that expanded the same 2K sentiment words into 175K words using a rule-based technique to match with variants, or forms, from that same corpus. We also utilise the cosine word similarity feature of the embeddings, not to retrieve words, but to associate the new words with sentiment scores and disambiguate them after the expansion.

In (Soricut and Och, 2015), the authors generated morphological inflections computationally by adding prefixes and suffixes to a set of vocabulary. They then measured the cosine similarity of each original word with the generated form to evaluate the morphological generation. Although this work is similar in concept to our work, the complexity of Arabic morphology is beyond simple prefixes and suffixes as mentioned in Section 1, hence we do not transform the vocabulary by adding affixes, rather we map the structure of that vocabulary with words of similar structure in the corpus. Similarly, we measure the cosine word similarity to rank the retrieved forms.

Other works considered the non-standard orthographies, such as the texts found on social media, a noisy or ill forms of the correct standard languages, thus emphasised their efforts on matching the equivalent forms of the noisy text with that of the standard orthography. For example, (Han and Baldwin, 2011) developed a classifier to detect the noisy forms, then generated correction candidate words based on the morphophonemic similarity. (Contractor et al., 2010) replaced the noisy words with a list of possible regular forms in a given set of sentences, then maximised the product of the weighted list and language model scores to generate cleaner versions of these sentences. (Gouws et al., 2011) used an unsupervised approach to extract word variants, in pairs, from domain specific corpora. They computed the distributional similarity among all word pairs in a domain specific corpus, filtered the common English ones, and scored the relation of the pairs using their semantic and lexical similarities. (Derczynski et al., 2013) also tried to automatically replace noisy words with standard English equivalents. Similar works presented in the (Baldwin et al., 2015) shared task on noisy user-generated text, where the participants were specifically asked to convert non-standard words to their standard forms for English tweets.

Most of these works considered that a correct form of the noisy words exist, therefore perceived the

task as finding the best word replacement, similar in concept to text simplification (Nisioi et al., 2017) which aims at transforming original words into simpler variants. This is similar to our problem except that there is no correct or wrong way to transcribe a spoken language that lacks a standard orthography, such as Arabizi; for that, our goal is to find all possible morphological and orthographic forms, for any given word, that are naturally written in text. The issue of non-standard orthography is not limited to Arabizi, but to many spoken languages and dialects (Millour and Fort, 2020), we therefore aspire that our contribution spans across other such languages as well.

3 Background and Challenges

Arabizi scripture contains several linguistic complexities that it is even called hell in (Seddah et al., 2020). We now scrutinise the transcription of Arabizi to explain how these complexities came to exist.

First of all, because dialectal Arabic (DA) is spoken, in its nature it lacks a consensus on an orthography, for that transcribing Arabic phonemes in a script of different languages (Latin) heuristically widens the orthographic inconsistency even further for the reasons summed below.

1. A word is transcribed based on how it is pronounced, hence different pronunciations of the same word spell differently. For example, the one letter vowel phoneme ي /yā/ in the positive word خير (*fine*) is pronounced as /āy/ khāyr or /eh/ kher, therefore both *khayr* and *kher* are common transcriptions for the same word.

2. There is no consistency in transcribing vowel phonemes. Both *khāyr* and *kher* could be transcribed as *kher, kheir, khair, kheyr, khayr,* or even without most vowel letters such as *khyr*.

3. There is little consistency in transcribing distinct consonant phonemes, such as the guttural ح ā', خ Khā', ع ayn, غ Ghayn that articulate in the post-velar areas of the oral cavity and the glottal stop ء Hamzah. For example, the خ Khā' in the mentioned word خير *khāyr* is standardised to some extent as compound letters *kh* or numeral 5 or even 7'. As such, all the following are possible orthographies for this tri-literal base word: *kher, kheir, khair, kheyr, khayr, khyr, 5er, 5eir, 5air, 5eyr, 5ayr, 5yr, 7'er, 7'eir, 7'air, 7'eyr, 7'ayr, 7'yr*. Note that different regions adopted different standards.

4. Finally, switching between Arabic and English or French is normal in DA speech. This is reflected in Arabizi texting as well, known as codeswitching (Aboelezz, 2009).

Additionally, transcribing a high number of phonemes into lower number of letters generates word ambiguity and causes Arabizi words to overlap with English words. However, in this work we focus on addressing the sparsity challenge.

This large lexical sparsity of Arabizi is even faced with a severe lack of resources. The existing corpus of 53MB Lebanese dialect Arabizi (Tobaili et al., 2019) is considered insufficient to train powerful data-hungry language models such as BERT and GPT (Devlin et al., 2018; Radford et al., 2018); a fundamental challenge for low-resourced languages.

On another note, attempting to convert Arabizi to Lebanese Arabic script, similar to the sophisticated works of (Darwish, 2014; Al-Badrashiny et al., 2014; Guellil et al., 2017; Masmoudi et al., 2019), serves us no purpose for lexical induction. We simply exploit the existing corpus to maximise the forms of given words using a rule-based combined with word embeddings approach.

4 Approach

We use the datasets that we previously created in (Tobaili et al., 2019) to test and evaluate the proposed approach. Briefly, we developed an Arabizi sentiment lexicon of 2K words, composed of 600 positive and 1.4K negative words separately, and expanded it to 24.6K words using the cosine word similarity feature of word embeddings trained on a corpus of 1M Facebook comments. We evaluated the lexicon using a simple lexicon-based approach against a sentiment-annotated dataset of 1.6K Arabizi tweets.

We now propose to expand that original sentiment lexicon (2K words) through several steps. We first normalise the text and search the entire Facebook corpus for all words matching the structure of the lexicon words under a linguistic constraint of permitted affix letters. We follow by adding cosine word similarity scores obtained from the word embeddings between every matched word and the original lexicon word to use it as a sentiment score and disambiguate forms that are retrieved in both positive and negative lists. We detail each step below.

1. **Normalisation:** Since transcribing consonant letters is standardised to a low extent in Arabizi as shown in Section 3, we normalise all observed compound letters that represent consonant phonemes with single numerals or symbols of a similar grapheme: (ch or sh ش with $), (gh غ with 8), and (kh خ with 5). The phonemes of ذ ظ , and ث are pronounced as *z* and *s* in the Lebanese dialect, hence normalised by nature. We also reduce repeated consonant letters down to one letter in exaggerated words: *habibbbi* to *habibi*.

   We apply this normalisation to determine which words to match only. If there is a match, we retrieve the word in its original un-normalised format.

2. **Matching:** We match the consonant letter sequence (CLS) of each lexicon word with all words containing the same sequence in the corpus: *fa5r* (*pride*) (f5r) matches with *fakhour, fakher, fa5er, fa5eir, fakhooorrr* after normalisation.

   If more consonant letters precede or follow the sequence, we check if these letters fall within a predefined set of observed prefixes (*2 l s w t b y m n*) and suffixes (*k l t n h w y*): *fa5r* (f5r) matches with all the following inflections *benfe5r, mntfekhar, fakhourin, fa5rna, fakhourah* because the prefixes *b, m, n, t* and suffixes *n, h* are permitted affixes.

   We minimise the error by matching the lexicon words that contain a CLS of length 3 or more, leaving out 126 (6%) short lexicon words.

3. **Scoring:** The CLS matching retrieves new words that are syntactically similar with the original words, therefore not all retrieved words are forms of these words; irrelevant words that have similar structure match as well: *hadame* (*humour*) matches with its form *mahdoum* (*humorous or cute*) and *tenhadam* (*get destroyed*). Such ambiguity exists because some letters represent two distinct Arabic phonemes, *d* for both ض and د in this case.

   FastText word embeddings projects word vectors based on their syntactic and semantic information (Bojanowski et al., 2017), we therefore utilise this feature to score all retrieved words based on their similarity with the original words. We give the original words a confidence score of 1, and sort all retrieved words from 1 (most relevant) to 0 (least relevant). We presume that the closer the retrieved words are to the original sentiment words the more likely they are to be forms of these words, as such we consider the similarity scores to be a measure of relatedness, consequently a measure of sentiment as well.

4. **Disambiguation:** Scoring the retrieved words made some automatic disambiguation possible. First, in case of a duplicate word or more per list, positive or negative, we keep the word with the highest score, since it resembles a high similarity with a lexicon word. Second, in case of words appearing in both the positive and negative lists, we also keep the ones with the higher scores in the list they pertain to: the positive *zalameh* (*bold man*) ambiguates with the negative *zalim* (*unjust*) because the letter *z* represents two distinct phonemes ز and ظ in Arabic. Both words share the same CLS (zlm), therefore the approach retrieves the same list of forms. After the disambiguation, the positive forms remain in the positive list and the negative forms in the negative list.

   The potential of this approach was apparent with many sentiment words reaching over 1K forms. We observed the retrieved words to find that most of them were inflectional and orthographic forms, except for some lexicon words that matched a high number of irrelevant forms: *wafa2* (*lucky* or *agree*) has a unique CLS of (wf2) matched 840 relevant forms, *3atel* (*useless*) on the other hand,

shares a CLS (3tl) with a highly inflectional lexeme *b3atl* (*send to*), matched 1.36K words that are mostly irrelevant.

We acknowledge that perfecting the approach requires automatic filtering of the irrelevant matched words. We checked if the frequency of the matched words in the corpus or the number of words retrieved could serve as features to determine whether they are relevant to the original words, but unfortunately they do not. Both relevant and irrelevant words retrieved are syntactically similar, therefore filtering irrelevant words automatically might necessitates a measure of contextual similarity, which is what the more recent language models are capable of computing (Peters et al., 2018; Devlin et al., 2018; Radford et al., 2018). However, under the current circumstance of low-resource NLP, we were not able to train such models on the corpus at hand. Therefore, we added a manual filtering step.

5. **Filtering:** After expanding and disambiguating each word automatically, a Lebanese native post-graduate student checks whether the retrieved words, collectively[2], are mostly relevant or completely irrelevant to determine whether to expand the lexicon word. The student filtered out 103 (5%) lexicon words; although small in number, if each word expands to 1K, we harm the lexicon with 100K irrelevant words.

As a result, the proposed approach expanded 1.7K words to around 175K, a new morphologically and orthogaphic rich Lebanese Arabizi sentiment lexicon. We named it SenZi-Large and released it publicly on the project's page[3].

5 Evaluation

We now evaluate the proposed approach extrinsically by applying a lexicon-based sentiment classification against the small annotated dataset of 1.6K tweets mentioned in Section 4. We wish to have had a larger dataset to reveal the true coverage of SenZi-Large, nevertheless, we settle for showing the value of the lexical induction approach.

In the lexicon-based approach, we search every word from the annotated tweets across the sentiment lexicon to assign scores to the matched words. We set the associated score for words that match with the positive list, and set a negated associated score for words that match with the negative list. Finally, we sum the scores to classify the tweets as positive or negative.

Since this is a binary classification task, similar to Subtask B from SemEval-2017 Task 4 (Rosenthal et al., 2017), we present the percentage of the classified tweets and the classification results of these tweets separately in Table 1. We compare SenZi-Large (175K words) with the original lexicon (2K words) and the previous expansion SenZi (24.6K words) that is mentioned in Sections 2 and 4.

Table 1: Extrinsic Evaluation
Lexicon-Based Sentiment Classification
1.6K Tweets (800 positive, 800 negative)

Lexicon	Classified	R	P	F	A
Original	22%	0.94	0.87	0.90	0.90
SenZi	55%	0.94	0.77	0.85	0.82
SenZi-Large	68%	0.83	0.80	0.82	0.83

Recall, **P**recision, **F**1-Score (Macro-averaging), **A**ccuracy

Inducing morphological and orthographic forms for Arabizi using the proposed approach improved the lexicon coverage by a solid 13% over the previous expansion. The induction expanded 1.7K words to 175K, that is more than 100 new forms per lexicon word on average. Although we disambiguated the

[2]For efficiency, the student does not validate the retrieved words word by word, rather skims the whole list.

[3]`https://tahatobaili.github.io/project-rbz/`

retrieved words and filtered out some lexicon words, as expected, such a large induction introduced some noise, neutral or irrelevant words, that traded-off the the coverage gain for a 3% loss in the F1-Score over the previous version.

We now asses the classification errors to point out the limitations of the produced lexicon. Since the evaluation of the lexical induction was done on sentiment classification of Twitter data, we asses the error manually from Twitter words that were either unclassified or falsely classified.

We extract a sample of 100 tweets at random from the 1.6K tweets dataset, of which 58 were correctly classified, 12 falsely classified, and 30 unclassified. We balance the 12 falsely classified with 12 unclassified tweets at random for the error analysis. From these 24 tweets, we counted a vocabulary of 178 words[4] that contain 30 sentiment words. The approach missed 24 sentiment words and falsely classified 19 non-sentiment words. We present the distribution of error for each case below in Table 2.

Table 2: Error Analysis

Sample	Words	Error	Distribution
Unclassified Sentiment Words	24	Orthographic Variant	46%
		Code-switched	37%
		Other	17%
Falsely Classified Non-Sentiment Words	19	False Neighbour	42%
		Other	58%

Most of the sentiment words that were not covered by the lexicon are orthographic variants of words that do exist in the lexicon: SenZi contains *bitwatar, bietwatar, btwaterne, betwaterne...* forms of (*I get nervous*) but not *betawattar* that appeared in text. Codeswitching also appears to be an inevitable challenge in Lebanese Arabizi texting: *chu heydaa ktiir cute* (*whats that* (expression) *soo cute*). On the other hand, most of the falsely classified non-sentiment words are false neighbours that were retrieved in the induction: *m3alim* (*an expert*) matched with a false neighbour *ma3loumeh* (*piece of information*). The other types of unclassified or falsely classified words consist mainly of multi-word sentiment expressions and sarcasm or contextual and polysemic words, which are known challenges for the lexicon-based sentiment analysis in general (Liu, 2012).

6 Discussion

Spoken languages that lack a standard orthography tend to have a high degree of lexical sparsity, especially if they are morphologically rich as well. If words can be transcribed in a variety of ways without a consensus on an orthography, then they can not be referenced with one correct orthography. This by nature defies word classification tasks such as sentiment analysis, NER, POS tagging, word sense disambiguation (WSD), anaphora resolution, etc.

We proposed a new approach to match words with their inflectional and orthographic forms in a low-resourced context. We hand-crafted a rule based approach that worked well with the linguistic nitty gritty complexities of a highly-sparsed scripture, Arabizi, and coupled it with word embeddings using minimal resources. The limitation of this approach was adding a manual check to filter out the lexicon words that retrieved irrelevant words. In a resourceful situation however, we would have trained advanced word embedding models to measure the contextual similarity of the words, with the intention to automate this step. Such models require large amount of data to pre-train especially for uncommon languages like Arabizi. Nevertheless, the result was a new sentiment lexicon, SenZi-Large, composed of 175K forms induced from 1.7K words, 7 times larger than the previous expansion for a 3% trade-off in F1 score. We presented a word-level classification error analysis on the same Twitter data to learn that most of the missed words are either orthographic variants of SenZi's words or English sentiment words and most of the falsely classified words were false neighbours introduced in the large induction. This consolidates the significance of lexical sparsity in non-standard languages.

[4]Excluding short words that consist of one or two characters only and non-alphanumeric words.

Although the outcome is a new public resource, the main contribution of this work is the proposed approach. The essence of the approach is in normalising and connecting consonant letters which comprise the backbone of most Latin script vocabulary. In the case of Lebanese dialect Arabizi, we observed a set of prefixes and suffixes and used them to minimise the error in the induction. However, given the complexity of Arabic morphology described in Section 1, a rule-based approach is far from success. We therefore combined it with word embeddings to filter the irrelevant and score the sentiment words. As such, our intuition is that the approach could potentially be useful for at least other Arabizi dialects if not other Latin script languages that pose similar challenges, such as Javanese dialects and Alsatian (Millour and Fort, 2020). The approach might also be catered for other NLP tasks such as lemmatisation or text simplification in morphologically-rich languages.

7 Conclusion

In this work we addressed the issue of high-degree lexical sparsity for a social language under a severe circumstance of small resources that are considered insufficient to train recent powerful language models. We proposed a new rule-based approach that utilises classical word embeddings to connect words with their inflectional and orthographic forms from a given corpus. Our case example is the low-resourced Lebanese dialect Arabizi, hence we used and expanded the only available resources we found in the literature to test and evaluate the approach. We induced 175K forms from a list of 1.7K sentiment words. We evaluated this induction extrinsically on a sentiment-annotated dataset pushing its coverage by 13% over the previous version. We named the new lexicon SenZi-Large and released it publicly[5].

Latinising a non-Latin script language heuristically is not a novel phenomenon, as to Arabizi, Greek, Farsi, Hindi, Bengali, Urdu, Telugu, Tamil and other languages are also transcribed in Latin script, known as Greeklish (Androutsopoulos, 2012), Finglish, Hinglish (Sailaja, 2011), Binglish, etc. For a future work, we plan to explore cross-lingual approaches (Ruder et al., 2019) to bridge such low-resourced languages that are written in different scripts.

References

Mariam Aboelezz. 2009. Latinised arabic and connections to bilingual ability. In *Lancaster University Postgraduate Conference in Linguistics & Language Teaching. Lancaster, UK*, volume 3, pages 1–23.

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.

Jannis Androutsopoulos. 2012. Greeklish": Transliteration practice and discourse in the context of computer-mediated digraphia. *Orthography as social action: Scripts, spelling, identity and power*, pages 359–392.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

Ramy Baly, Hazem Hajj, Nizar Habash, Khaled Bashir Shaban, and Wassim El-Hajj. 2017. A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–21.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Danish Contractor, Tanveer A Faruquie, and L Venkata Subramaniam. 2010. Unsupervised cleansing of noisy text. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 189–196. Association for Computational Linguistics.

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

[5] https://tahatobaili.github.io/project-rbz/

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Stephan Gouws, Dirk Hovy, and Donald Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90. Association for Computational Linguistics.

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*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 368–378.

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

Abir Masmoudi, Mariem Ellouze Khmekhem, Mourad Khrouf, and Lamia Hadrich Belguith. 2019. Transliteration of arabizi into arabic script for tunisian dialect. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):1–21.

Alice Millour and Karën Fort. 2020. Text corpora and the challenge of newly written languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 111–120.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Pingali Sailaja. 2011. Hinglish: code-switching in indian english. *ELT journal*, 65(4):473–480.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content north-african arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150.

Radu Soricut and Franz Josef Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637.

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

Taha Tobaili, Miriam Fernandez, Harith Alani, Sanaa Sharafeddine, Hazem Hajj, and Goran Glavaš. 2019. Senzi: A sentiment analysis lexicon for the latinised arabic (arabizi). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1203–1211.

Taha Tobaili. 2016. Arabizi identification in twitter data. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 51–57.

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