4

A Deep Learning Approach for Sentiment and Emotional Analysis of Lebanese Arabizi Twitter Data

Maria Raïdy and Haidar Harmanani

Abstract

Arabizi is an Arabic dialect that is represented in Latin transliteration and is commonly used in social media and other informal settings. This work addresses the problem of *Arabizi* text identification and emotional analysis based on Lebanese dialect. The work starts with the extraction and construction of a dataset and uses two machine learning models. The first is based on *fastText* for learning the embeddings while the second uses a combination of recurrent and dense deep learning models. The proposed approaches were attempted on the *Arabizi* dataset that we extracted and curated from Twitter. We attempted our results with six classical machine learning approaches using separate sentiment and emotion analysis. We achieved the highest result in literature for the binary sentiment analysis with an F1 score of 81%. We also present baseline results for the 3-class sentiment classification of *Arabizi* tweets with an F1 score of 64%, and for emotion classification of Arabizi tweets with an f1 score of 61%.

Keywords

Emotional Analysis · Arabizi · Deep Learning

4.1 Introduction

Social media has been at the center of the digital age causing disruptions from the rise of political polarization to the COVID-19 pandemic. According to the *Pew Research Center*, seven-in-ten Americans use social media to connect with one another, engage with news content, share information, and express their happiness, frustrations, and anger [1]. Various countries have resorted to social media in order to measure the wellness of their own citizens, not an easy task given that such measures include evolving economical, environmental, and social indicators.

Some countries regularly mine social media in order to report welfare statistics and use them to detect social anxiety or sudden decrease in satisfaction. For example, the United Kingdom looks at societal and personal well-being through areas such as health, relationships, education, skills, finance, and the environment [2]. Although there are no government-led efforts in Lebanon to study societal well-being, various researchers looked at Lebanese well-being in a variety of psychological contexts [3–10]. Other researchers examined the impact of plurality and communitarianism on well-being [11–13] in addition to the Lebanese social media habits during such events such as the *Beirut Port Blast* [14–20] or during social unrest [21–23]. During the 2019 social unrest and the economical meltdown of 2020, the New Economics Foundation (NEF) ranked Lebanon 120th with a *Happy Planet Index* (HPI) score of 21.9 [24].

This paper uses a deep learning approach in order to analyze sentiment and emotions of Lebanese Arabizi tweets during the 2019 and the economical meltdown of 2020. We create, curate, and label a Lebanese Arabizi dataset and then compare our results with six classical machine learning approaches using separate sentiment and emotion analysis tasks separately. Finally, we use these results in order to determine the Lebanese Social Happiness Index (LSHI) during the same period.

4.2 Arabizi

Arabizi is a word that is composed of “Arabic” and “englizi,” the Arabic word for English. It is an Arabic dialect that is

M. Raïdy · H. Harmanani (✉)
Department of Computer Science and Mathematics, Lebanese American University, Byblos, Lebanon
e-mail: haidar.harmanani@lau.edu.lb

S. Latifi (ed.), *ITNG 2023 20th International Conference on Information Technology-New Generations*, Advances in Intelligent Systems and Computing 1445,
https://doi.org/10.1007/978-3-031-28332-1_4

represented in Latin transliteration and used increasingly by millennials as it makes writing what's being spoken easier [25–27]. Arabizi is easy to use, fast to type, and lacks complex grammatical or syntax rules. In fact, it is the language of choice for younger bilingual Arab generations who *code switch* between languages on social media as well as in texting or emailing.

In multilingual countries such as Algeria, Tunisia, and Lebanon, Arabizi has evolved to include French. In an extensive study of Arabizi, Sullivan [28] noted that Lebanese Arabizi is more complex than other Arabizis as most Lebanese are trilingual and tend to use several languages (Arabic, English, French, and/or Armenian) in the same sentence. The authors illustrates this case using the classical Lebanese salutation: "Hi! kifak? cava?"[1] which mixes Lebanese Arabic (Kifak), French (Ça va), and English (Hi) in the same sentence. Note that in this case the French *cédille* is omitted from "Ça va."

There are various Arabizi versions depending on the local Arabic dialect. Boudad et al. [29] noted that social media is found to include 30 different Arabizi versions. These informal texts lend themselves to ambiguity as they disregard grammar, capitalization, and punctuation rules. There are no *formal* rules on how to mix Arabic and Latin characters in Arabizi. However, in most variations, Latin characters and numbers are used in order to represent sounds that do not exist in such languages. For example, the guttural Arabic letter ع or *ayn* is typically represented using "3", "aa" or even "3a" while the voiceless fricative ح or Haa is typically represented using "7." Some variations may completely omit all vowels for simplicity.

4.3 Related Work

Although a lot of research investigated NLP techniques for sentiment analysis, emotion analysis, or translation, few tackled Arabizi due to the recent emergence of the dialect and its informality and complexity. Guellil et al. [30] and Duwairi et al. [31] used Naïve Bayes and SVM machine learning techniques in order to classify sentiments of Arabizi messages into positives or negatives. Both transliterated the local Arabizi dialect (Algerian and Jordanian, respectively) into formal Arabic first. Barr et al. [32] chained two distinct methodologies, word2vec and a graph-theoretical algorithm, in order to classify documents into distinct classes. Liu et al. [33] proposed an adaptive DNN inference acceleration framework that accelerates DNN inference. Srour et al. [34, 35] used a mixed theme/event based approach to rate highly influential users in Twitter. Hajjar et al. [36] and Sarkissian et al. [37] used an unsupervised approach to identify sentences having the highest potential to represent informative content in a document. Wang et al. [38] used a deep learning model that combined Conditional Random Field and Bidirectional LSTM for NER with unbalanced labels. Thomas et al. [39] explored the relative frequency of social attributes in Tweets across both English and Arabic. Abd El-Wahab et al. [40] explored trends in transliteration using deep learning models and propose an approach for Arabic-English transliteration. Darwish [41] used CRF sequence labeling in order to identify Arabizi in texts prior to converting it to formal Arabic. Bies et al. [42] created a corpus for the purpose of Egyptian Arabizi transliteration into Arabic. Tobaili [43] tackled sentiment analysis for Lebanese Arabizi using an SVM classifier. Kiritchenko et al. [44] detected sentiments behind short informal texts extracted from twitter and SMS using a lexicon-based approach. Kundi et al. [45] developed a tool to detect and analyze interesting slang words while Bahrainian et al. [46] created and used a slang dictionary for classification. Dhaoui et al. [47] tested a lexical sentiment analysis approach versus machine learning approaches.

4.4 Dataset Generation

Twitter-based textual data provide a trove of information that are domain specific and tackle various topics. Tweets are attractive since they are limited in length to a maximum of 280 characters, highly available through APIs, and easily accessible from different media. Since there is no publicly available Lebanese Arabizi dataset that can be used for sentiment and emotion analysis, we did curate, annotate, and label our own data set. We scraped data from Twitter based on a set of 30 common keywords in Lebanese, such as "*lech, chou, hayda, kamen, hasset, chwei, ...*" The keywords served as seeds to collect tweets published between 01-01-2017 and 31-04-2020. The data was also geographically filtered based on the Tweet and account location. We excluded all spams, images, noise, duplicates, retweets, tweets containing Arabic characters, and tweets that only consist of hashtags and/or URLs. The initial data set included 135,000 Lebanese Arabizi tweets written in English, French, and Armenian.

The dataset was next manually inspected and verified. Tweets that have at least 50% Arabizi words were kept, resulting with 46,000 Arabizi tweets. Next, two independent annotators performed sentiment and emotional labeling on the dataset and only kept the tweets the two annotators agreed on. The labels varied between *polarity* (positive, negative, or neutral) and *emotion* according to the Ekman model (*happiness, anger, sadness, fear, surprise, and disgust*). Finally, we added an additional set of social labels that are of particular interest in Lebanon such as *sectarianism, gossip, sexism,*

[1] hello! How are you? Is all good?

Table 4.1 Examples of sentiment classifications in our dataset

Polarity	Emotion	Arabizi tweet	Translation
Positive	Joy	Khay Shou helo	Oh, how beautiful!
		habbet hal jemle !:p	I liked this sentence!
		nzale btetsalle ktir :p	Go! You'll enjoy it!
Negative	Sadness	Ana kamen bkit bas ma habbet abadan	I also cried but I didn't like it at all
		Kteir z3elet.. :(	It made me so sad
	Bullying	Lea ktir pedophile 😂	Lea is too much of a pedophile
	Gossip	Aade manna kitr "waw" ya3ne. Elle e troo aade	She's not that pretty. She's average.
	Fear	aywaa aywaa nshalla nnja7 sara7a ana ktir khayfe :/	I hope we succeed, I'm honestly so worried.

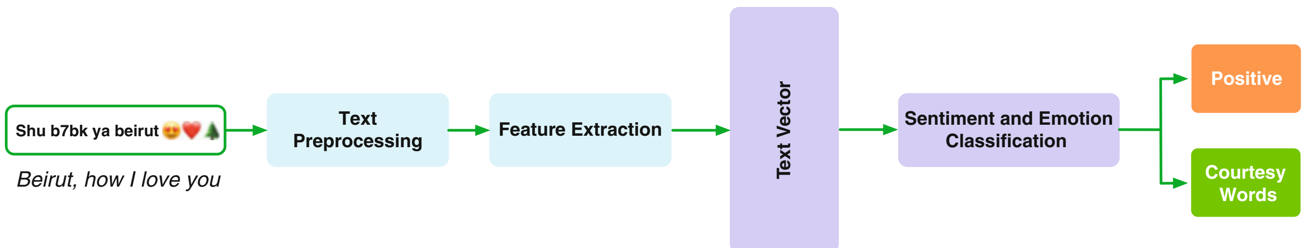

Fig. 4.1 Classification system architecture

racism, bullying, sarcasm, and *foul language*. A sample set of tweets and corresponding labels are shown in Table 4.1. The dataset is available on Kaggle.

4.5 Classification System

We analyzed and classified sentiment and emotions of Lebanese Arabizi tweets and then compared our results with six classical machine learning approaches. In what follows, we describe our classification system with reference to Fig. 4.1.

4.5.1 Text Pre-processing

Preprocessing is an important step in our approach because of the informality of the Arabizi language and the multiple typos and errors that are the byproducts of social media. During this step, we convert all the characters to lowercase, and remove user mentions, URLs, special characters, numbers, measurements, and timings. We do, however, keep all the words that might give an insight into the polarity and emotion of the tweep. We also simplified the texts exaggerations and deleted all stop-words using a list of 398 tokens specific to our data. The stop-words list included a combination of English and French words as well as the most common Arabizi terms found in our corpus. Finally, we cleaned the dataset from empty tweets and balanced it using random over sampling. The result was two balanced datasets of 1800 tweets for sentiment analysis and 1500 for emotion analysis.

Table 4.2 Sound-effects tags with examples

Tag	Examples
@laughter	hahahahhahaaah, lololool, hihihihihihi, wahaha
@amazement	yaayyyy, wooowow, wouw
@surprise	ohhh, aaahh, owww
@annoyance	mehhh, ufft, ughh, huh, hush, hushtt
@thinking	uhmm, ummm, mmm, heinn
@kiss	mwaahhh, mmmmwh
@sound-effect	uff, pshht, psst, shhhht, a777

Table 4.3 Emojis and emoticons

Tag	Examples
:-) :) :-] :3 :8-) :-} =)	happy_face_smiley
:-(:(:c ::-[:-| :@	frown_sad_andry_or_pouting
:-* :* :X	kiss
👍	thumbs_up

4.5.2 Feature Extraction

The next step was analyzing the dataset for possible features extraction. In this stage, we kept negating words (la2/la /no/, ma/mich /not/, ...) to preserve the purpose of the tweet. We also tampered down the enthusiasm in the tweets and used different tags to replace exaggerations and sound effects such as @laughter, @amazement, @surprise, @annoyance, @thinking, @kiss and/or simply @sound-e ect (Table 4.2). We also replaced the emojis and emoticons by their description (Table 4.3). All symbols and sounds were normalized in order to account for different regional local dialects (Table 4.4).

Table 4.4 Normalized representation of complex sounds

Arabic Script	Phonetic Alphabet	Common Arabizi Representations	Normalized Representation
ح	ʒ	j/g	j
ج	ħ	7/h	h
ح	x	7'/5/kh	kh
خ	ð	d/z	Can't normalize
ذ	ʃ	sh/ch	ch
ش	ðˤ	th/z	Can't normalize
ظ	ʕ	3/aa	aa
ع	ɣ	3'/8/gh	gh
غ	q	9/q	q
ق	w/u	o/w/ou/u	u
و	y/i	i/y/ey/ei	i

4.6 Sentiment and Emotion Classification

Dictionary-based approaches are not efficient when dealing with an informal, morphologically rich, and constantly evolving language like Arabizi. In order to classify the Arabizi text, we used multiple models that we will describe next.

4.6.1 Text Vectorization Using Fasttext

We used the skip-gram word embeddings model resulting with a $V \times D$ matrix where V is the vocabulary size and D is the embeddings dimension [48]. One of the limitations of morphologically rich languages with large vocabularies and many rare words such as Arabizi is that this model assigns a distinct vector to each word ignoring its morphology. To resolve this problem, we use the fasttext library which includes subword information [49]. Each word vector is a sum of the vectors of its character n-grams. As a result, they deal with out-of-vocabulary words, such as misspelled words, by being able to embody them by their sub word representations. We improve the model on our dataset.

We trained the system on our labeled data in order to predict if a tweet is in Arabizi or not. We used a learning rate of 0.05 and 3 epochs. We set the word n-gram window size was to 3, and based our training on sub word features with character n-grams of size 2–6. We also used hierarchical softmax activation function. We obtained 95% precision recall and F1 following fivefold validation and 98% precision recall and F1 score on the *arabizi-detection-validation-set*. The results are the highest in reported Arabizi classification.

In order to assess the impact of the text pre-processing stage, we repeated the above experiment using the same model and parameters on the raw unprocessed dataset. We received 94% precision recall and F1 score on the balanced test set and 94% precision 93% recall and F1 score on the Arabizi detection validation set. Hence, the preprocessing method improved our results by 4%. Finally, combining both methods using the automatically Arabizi classified tweets with a ratio of 50% to train our model resulted with an F1 score of 96% on testing and 97% on validation.

4.6.2 Machine Learning Approaches

We next attempted six classical machine learning models including the Multinomial Naive Bayes, Random Forest, SVM (with linear, Gaussian, and sigmoid kernels), decision trees, logistic regression, and a vanilla Neural Networks model. The neural networks model was based on two hidden layers of size 100, the 'Adam' optimizer, 300 epochs, and warm start turned on. Each model was ran 768 times with different choices at the different stages.

4.7 Deep Learning Model

The last model we attempted was a deep learning model that uses a combination of embedding layer, bidirectional LSTM recurrent networks, dropout layers, and dense layers. The model, shown in Fig. 4.2, used two LSTM layers. The first LSTM layer used 400 units while the second used 200. We also used three dense layers with the first having 400 units and the second 100. The last layer was the prediction layer and the number of units depended on the prediction label. In the polarity case it used a sigmoid unit while in the emotional analysis it used a softmax activation function.

4.8 Experimental Results

We present separate results of our finalized models for sentiment and emotion analysis in Tables 4.5 and 4.6, respectively. For sentiment analysis, the model with the best accuracy of 65% is the Logistic Regression approach with a count vectorizer of trigrams, tagged sound effects, handled negation, replaced emojis and emoticons, and normalized phonetics. Notice that we did not remove stop words in this case. The same model ran on an unprocessed dataset gets an accuracy of 63%. The downside of our approach is that we used all our data for training and testing since it's a rather small dataset and therefore did not have a validation-set to validate the accuracy of our models on unseen data. To be able to compare our results to those in the literature we trained the best models obtained for our 3-class sentiment analysis (Positive, Negative, Neutral) combined with the preprocessing methods

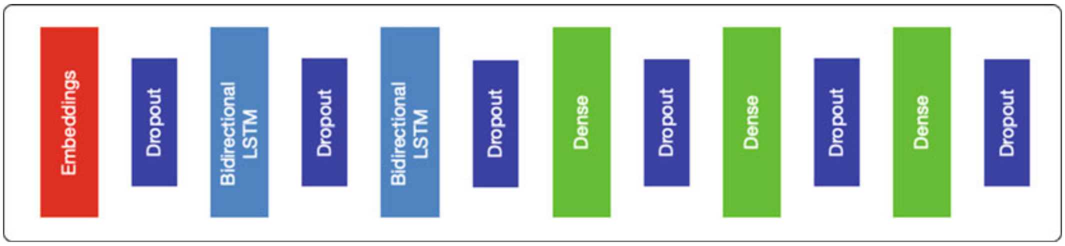

Fig. 4.2 Deep learning model architecture

Table 4.5 Sentiment analysis results for machine learning models

Algorithm	Accuracy	Precision	Recall	F1
Multinomial naive bayes	0.61	0.63	0.61	0.61
Random forest	0.60	0.63	0.60	0.60
SVM linear	0.63	0.64	0.63	0.63
SVM Gaussian	0.62	0.63	0.62	0.62
SVM sigmoid	0.62	0.63	0.62	0.62
Decision trees	0.56	0.56	0.56	0.55
Logistic regression	0.65	0.65	0.65	0.64

Table 4.6 Emotion analysis results for each model

Algorithm	Accuracy	Precision	Recall	F1
Multinomial naive bayes	0.58	0.58	0.58	0.58
Random forest	0.60	0.62	0.60	0.60
SVM linear	0.60	0.61	0.60	0.60
SVM Gaussian	0.59	0.61	0.59	0.60
SVM sigmoid	0.59	0.62	0.59	0.60
Decision trees	0.54	0.58	0.54	0.54
Logistic regression	0.60	0.62	0.61	0.61

opted by each on a 2-class sentiment analysis task (Positive, Negative). The best accuracy obtained is 81% with a decision tree model joined with tagged sound effects, removed stop words, handled negations, and normalized phonetics. As for the best performing model, it was the deep learning model shown in Fig. 4.2 and which achieved a validation accuracy of 68% as shown in Fig. 4.3.

One of the applications of our study is to measure the Lebanese Social Happiness Index (LSHI) based on Arabizi Twitter data. The sample represents 25% of the population who use Twitter frequently, and 33% of those who use Arabizi when on social media. It is interesting to note how well the tweets mood correlated with the events that occurred in Lebanon between the start of the October 17 revolution as well as with the start of the COVID-19 pandemic. Figures 4.4 and 4.5 show the results of sentiment and emotion analysis of our best models. Notice that if we align the above results with the COVID-19 updates in Lebanon, we notice an increase in the number of negative tweets whenever a pessimistic event had happened. For example, on the 21st of February 2020, when the first COVID-19 case was announced in Lebanon, we see a clear peek in anger and fear tweets. This is also the case when the first death case is announced followed by the first complete lockdown starting on the 10th of March 2020. However, anger, fear, and sadness dominated as the number of COVID cases increases as of May 2020.

4.9 Conclusion

We have presented a deep learning approach for Arabizi text identification and emotional analysis. The proposed approaches were attempted using six classical machine learning approaches. We achieved the highest result in literature for the binary sentiment analysis. We have also presented baseline results for the 3-class sentiment and emotional classification of *Arabizi* tweets. The reported results and comparisons were favorable (Fig. 4.6).

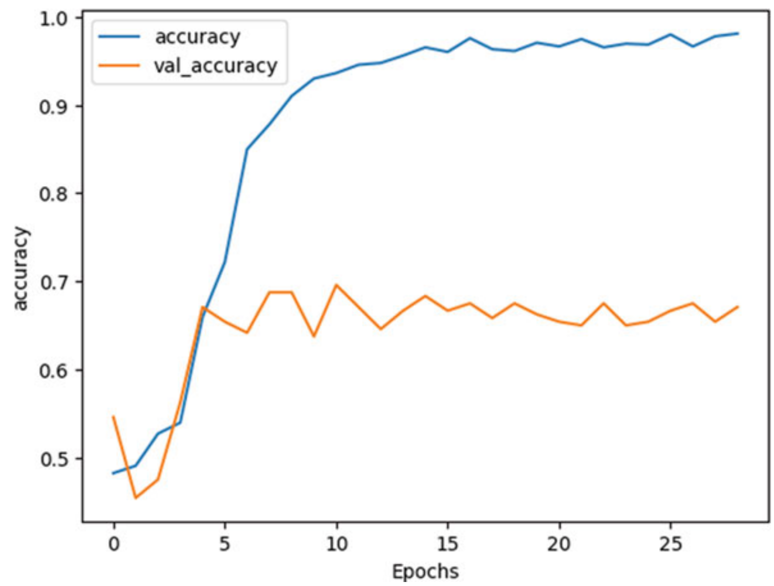

Fig. 4.3 Deep learning model validation and training accuracy

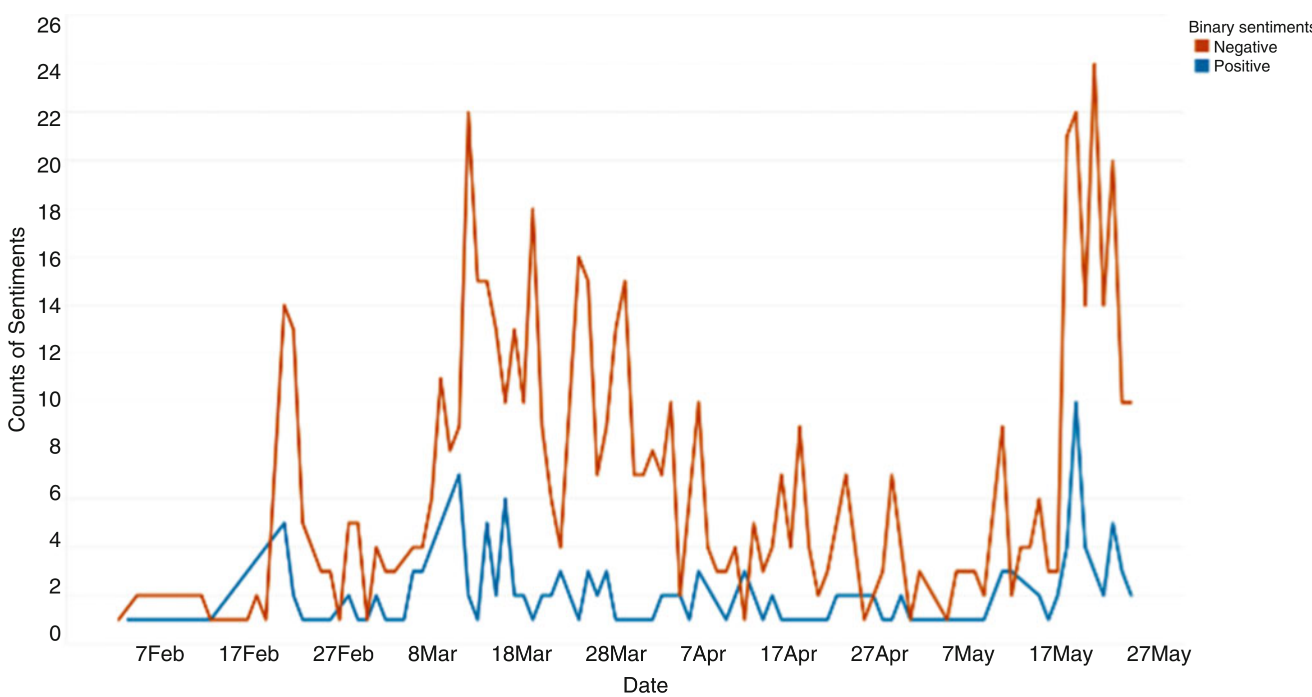

Fig. 4.4 Binary sentiments detected by the DT model in COVID-19 tweets between February 2020 and May 2020 (Lebanon) with 65% accuracy

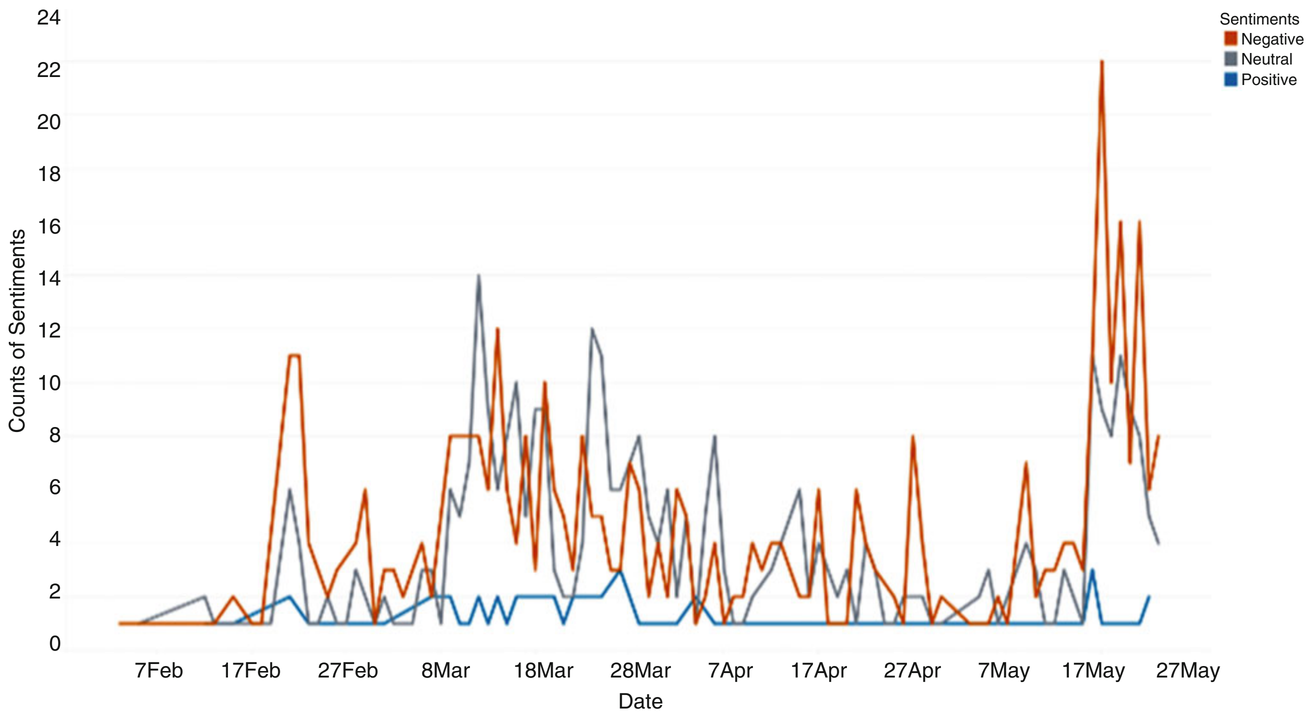

Fig. 4.5 Sentiments detected by the DT model in COVID-19 tweets between February 2020 and May 2020 (Lebanon) with 65% accuracy

References

1. Pew Research Center, Social Media Fact Sheet (Pew Research Center, Washington, 2021)
2. UK Office of Online Statistics, Well-Being (UK Office of Online Statistics, Newport, 2022)
3. P. Tohme, R. Abi-Habib, E. Nassar, N. Hamed, G. Abou-Ghannam, G. Chalouhi, The psychological impact of the covid-19 outbreak on pregnancy and mother-infant prenatal bonding. J. Child Health **26**(11) (2022)
4. P. Tohme, I. Grey, M. El-Tawil, M. El Maouch, R. Abi-Habib, Prevalence and correlates of mental health difficulties following the beirut port explosion: the roles of mentalizing and resilience. Psychol. Trauma Theory Res. Pract. Policy (2022). http://dx.doi.org/10.1037/tra0001328
5. R. Abi-Habib, R. Chaaya, N. Yaktine, M. Maouch, P. Tohme, Predictors of the impostor phenomenon. Asian J. Psychiatr. **75**, 1–2 (2022)
6. M. Greaves, The reflexive self; adapting to academic life in a time of social turmoil. Teach. High. Educ. 1–16 (2021)
7. D. Nauffal, J. Nader, Organizational cultures of higher education institutions operating amid turbulence and an unstable environment: the lebanese case. High. Educ. **84**, 343–371 (2022)
8. A. Al-Shehhi, I. Grey, J. Thomas, Big data and wellbeing in the arab world, in *Positive Psychology in the Middle East/North Africa: Research, Policy, and Practise*, ed. by L. Lambert, N. Pasha-Zaidi (Springer, Berlin, 2019), pp. 159–182
9. N.B. Zakhem, P. Farmanesh, P. Zargar, A. Kassar, Wellbeing during a pandemic: an empirical research examining autonomy, work-family conflict and informational support among SME employees. Front. Psychol. **13** (2022)
10. R. Gök, E. Bouri, E. Gemici, Can Twitter-based economic uncertainty predict safe-haven assets under all market conditions and investment horizons? Technol. Forecast. Soc. Chang. **185**, 1–21 (2022)
11. A. Desatnik, C. Jarvis, N. Hickin, L. Taylor, D. Trevatt, P. Tohme, N. Lorenzini, Preliminary real-world evaluation of an intervention for parents of adolescents: the open door approach to parenting teenagers (apt). J. Child Fam. Stud. **30**, 38–50 (2021)
12. I. Salamey, The communitarian nation, in *The Communitarian Nation-State Paradox in Lebanon*, ed. by I. Salamey, chap. 10 (Nova Science Publishers, Hauppauge, 2021), pp. 266–290
13. I. Salamey, Reconstructing the communitarian state, in *The Communitarian Nation-State Paradox in Lebanon*, ed. by I. Salamey, chap. 13 (Nova Science Publishers, Hauppauge, 2021), pp. 309–333
14. S.E. Hajj, Archiving the political, narrating the personal: the year in lebanon. Biography **42**, 84–91 (2019)
15. S.E. Hajj, Golgotha, beirut: a feminist memoir of the port blast. J. Int. Women's Stud. **24**, 1–4 (2022)
16. G. King, Radical media education practices from social movement media, in *The Handbook of Media Education Research*, ed. by D. Frau-Meigs, S. Kotilainen, M. Pathak-Shelat, M. Hoechsmann, S.R. Poyntz, chap. 40 (Nova Science Publishers, Hauppauge, 2020)
17. C. Kozman, R. Cozma, Keeping the gates on twitter: interactivity and sourcing habits of lebanese traditional media. Int. J. Commun. **15**, 1000–1020 (2021)
18. C.-J. El-Zouki, A. Chahine, M. Mhanna, S. Obeid, S. Hallit, Rate and correlates of post-traumatic stress disorder (ptsd) following the beirut blast and the economic crisis among lebanese university students: a cross-sectional study. BMC Psychiatry **22**(1), 532 (2022)
19. G. Sadaka, Illnesses of illusion and disillusionment: from euphoria to aporia. Life Writing J. 1–12 (2022). Published Online
20. P. M. B. Doleh, Truman in Beirut: Journeying through fear and immobility," *Life Writing*, pp. 1–13, 2022.
21. J.G. Karam, Teaching political science during crisis: The three-c approach and reflections from lebanon during a social uprising, an economic meltdown, and the covid-19 pandemic. J. Polit. Sci. Educ. **18**, 492–510 (2022)

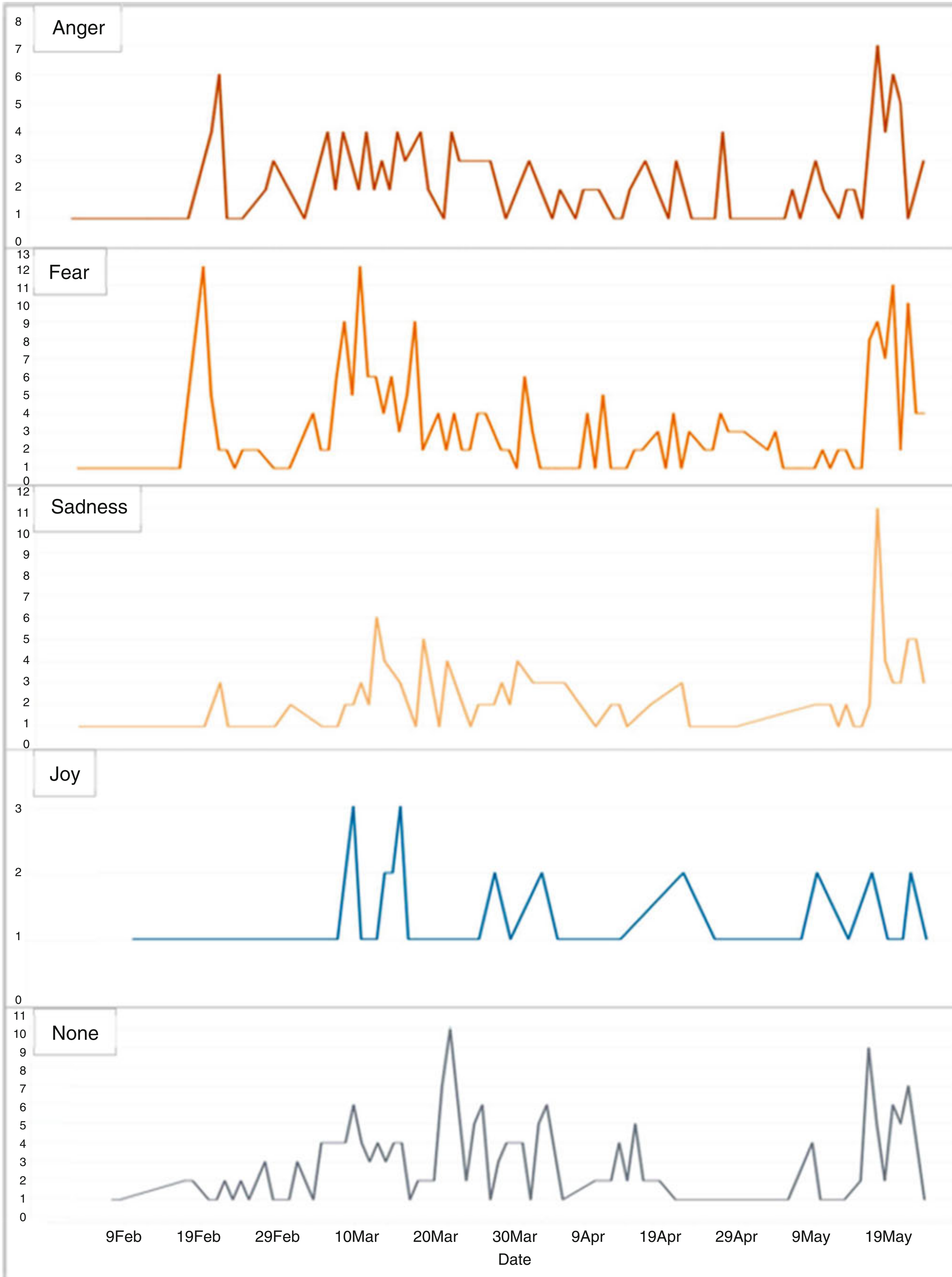

Fig. 4.6 Emotions detected by the LR model in COVID-19 tweets between February 2020 and May 2020 (Lebanon) with 61% accuracy