Algerian Arabizi rumour detection based on morphosyntactic analysis

Chahnez Zakaria, Kamel Smaïli, Besma Sahnoun, Assia Chala, Radjaa Agagna, Célia Amirat

▶ **To cite this version:**

Chahnez Zakaria, Kamel Smaïli, Besma Sahnoun, Assia Chala, Radjaa Agagna, et al.. Algerian Arabizi rumour detection based on morphosyntactic analysis. International Journal of Knowledge Engineering and Data Mining, 2023, 8 (1), pp.43-66. 10.1504/IJKEDM.2023.135716 . hal-04388759

HAL Id: hal-04388759
https://hal.science/hal-04388759

Submitted on 11 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rumor detection approach in Agerian Arabizi based on morphosyntaxic analyser

Chahnez ZAKARIA, Kamel SMAILI, Besma SAHNOUN, Radjaa AGAGNA, Assia CHALA, and Celia AMIRAT

[1] F. Author first address
Tel.: +123-45-678910
Fax: +123-45-678910
fauthor@example.com
[2] S. Author second address

Abstract. Social networks have become a customary news media source in recent times. However the openness and unrestricted way sharing the information on social networks fosters spreading rumors which may cause severe damages economically, socially, etc. Motivated by this, our paper focuses on the rumor detection problem in Algerian Arabizi. Studying linguistic rules of the Algerian Arabizi, we propose an lemmatiser and a parser for analysing and standardizing the text in order to produce better rumor detection models. An approach for classifying rumors and news in social networks based on the expression of emotions and positions of users is proposed. The experiments were done on many ngram representation and the best one has reached more than 94% for f-score. In addition to that this research deals with resources creation for Algerian Arabizi which is an under-resourced dialect. A corpus and several lexicons have been built and which can be the subject of other works dealing with this dialect.

Keywords: Social networks · Rumor detection · Lemmatizer · Parser · Arabizi

1 Introduction

With the expansion of the Internet and information technologies, social networks are becoming the main means of communication. The social networks are used to disseminate information and keep up with the latest news. These communication media allow a massive and rapid dissemination of information but at some risk. Not all shared information is accurate which allows the spread of rumors. Some rumors can start as trivial facts and grow over time to become a matter of public opinion. This is due to the fact that some Internet users quickly share content without verifying its veracity. It is therefore essential to detect these rumors to stop their dissemination.

Many websites like HoaxBuster[3], Snopes[4], Hoaxkiller[5] have been created to limit the spread of rumors. After identifying eventual rumors usually by the editorial staff of these websites sometimes with the help of the users, research is carried out in order to disentangle the true from the false information by contacting people or institutions that can validate or deny the information. Even if these websites could constrain the spread of rumors, they rely on a laborious working process which is often expensive and time consuming. Automatic rumor detection proves more than necessary especially regarding the speed of rumor propagation on social networks and their disastrous consequences.

Several works have undertaken rumor detection automatically, but since rumors have no specific format and could be governed by no rules of expression, the task proves so challenging. They're all trying to find the features that distinguish rumor from news. Various factors have been targeted as the user credibility information Li et al. (2019), the extent of propagation on the network Tu et al. (2021), the textual feature extraction Khanam et al. (2021), the social network feature extraction Qazvinian et al. (2011) Castillo et al. (2011), etc. In this paper, we aim to use only the text to differentiate between rumors and news, however we investigate in words that express the emotions users and their positions toward an event whether it is a news or a rumor. A considerable research investigating rumor detection have been undertaken in many languages Hamidian and Diab (2019) Guibon et al. (2019) Pathak et al. (2020) and a significant number in Modern Standard Arabic (MSA) Nagoudi et al. (2020) Al-Yahya et al. (2021) Mahlous and Al-Laith (2021). To the best of our knowledge there is one work on the detection of rumors on Algerian dialect written in Arabic script Alkhair et al. (2019), however no work do it on Algerian dialect written in Latin script (Known as Arabizi). Arabizi makes detecting rumors even more difficult. Indeed processing Arabizi is frought with challenges due to the absence of standard writing rules. To solve this problem some works propose to transliterate the Arabizi to Arabic script (called Abjadia) Al-Badrashiny et al. (2014) Bies et al. (2014). In GUELLIL et al. (2017) authors propose to transliterate Arabizi to Abjadia for analysing sentiments in the Abjadia instead of the Arabizi. In this case, errors in sentiment analysis can be magnified by transliteration errors. It is reason why we want to study arbizi for understanding its linguistic rules and modeling them.

In Social networks, it seems that Algerians use the Latin script more than the Abjadia script. However work on the Arabizi is still lagging behind specially, natural language processing tools. To our knowledge, there is one work to propose a morphosyntactic analyser in Guellil and Azouaou (2016). However it does not cover all the grammatical rules of Algerian Arabizi. In this paper, the detection of rumors in the Algerian Arabizi requires addressing two objectives which in fact correspond to two main contributions. First, we automate the grammatical rules that allow lemmatizing and assigning parts of speech (POS) to words in

[3] https://www.hoaxbuster.com/
[4] https://www.snopes.com/
[5] http://www.hoaxkiller.fr/

Algerian Arabizi text. Second, we propose lexicons of the emotions and reactions of Internet users, in order to distinguish rumors from news.
The rest of the paper is organized as follows: Section 2 presents specificities of Algerian Arabizi transcription. Section 3 provides an overview of related work. Section 4 presents our rumor detection approach. In section 5, we describe our morphosyntactic analyzer for Algerian Arabizi. Section 6 presents corpus collected, lexicons and vectorial representation schemes built for detecting rumors and news. In section 7, we present our experimentations and results. We conclude in Section 8.

2 Specificities of Algerian Arabizi transcription

2.1 Specificities social networks

In social networks, Algerians use two *transcription systems* - Latin and Abjadia - and several languages, namely French, MSA, Algerian dialect, English, etc. However they use more the Algerian dialect with the Latin transcription (i.e. the Algerian Arabizi). Abidi and Smaili (2017) presents a study made on a corpus of discussions between Algerians, collected from Youtube. It shows that 51% of the words of the vocabulary used, are in Latin transcription. This vocabulary includes words in MSA, French, English, Algerian Arabizi and Algerian dialect in Abjadia transcription. In the 51% of words in Latin transcription, except 2% and 3% correspond respectively to English and French. In other words, 47% of vocabulary words represents words in Algerian Arabizi.

The *code-switching* consists of alternating between several languages, possibly several times, in the same utterance [6] Joshi (1985). In social networks, Algerians combine several languages (MSA, dialect, French, English, etc.), in the same utterance. Indeed, in the corpus of Abidi and Smaili (2017), 82% of its comments are a mixture of two or more languages (MSA, dialect, French and sometimes English). The dialect comments represents only 9% of the total corpus.

2.2 Specificities phonological dimension

The specificity which makes the processing of Algerian Arabizi very difficult is the transcription diversity of the same term. Indeed, the Algerian Arabizi is sorely lacking in standards. This prompted speakers to express themselves spontaneously in social networks. This diversity is observed through three phenomena:

1. The use of several letters or combinations of letters for the same sound. For example, the use of the two combinations, «sh»and «ch»for the same sound, as in the two variants «chaba»and «shaba»(beautiful).

[6] Term coined by E. Haugen

2. The use of letters or combinations of letters, to create graphemes for sounds that do not exist in the Latin alphabet, like the sound in Arabic corresponding to «ح»which is designated by «h».
3. The use of numbers to write sounds that do not exist in the Latin alphabet, like the number «7»which indicates the sound «ح».

2.3 Specificities morphological dimension

In Algerian Arabizi, words (verbs, adjectives, and nouns) undergo modifications due to the influence of the gender and the number of neighboring words. The verb conjugation in Algerian Arabizi, is done by adding suffixes and / or prefixes to a stem. For example the verb «med»(to give), in the past takes the ending «ina»with the first plural pronoun. In Algerian Arabzi the feminine gender exists even in the second singular person. The verb «med»becomes «med**it**»with the pronoun «you»and becomes «med**iti**»in the feminine form. However the pronoun «It»does not exist. The tenses frequently used in Algerian Arabizi are the present, the past, the imperative and the future. The future conjugation is done in the same way as that of the present, however it is recognized through the use of time markers such as the adverbs of time «ghedwa»(tomorrow).

Generally, we modify the masculine of the adjective to the feminine by adding the suffix «a», the masculine plural by adding «in»and the feminine plural by adding «at». For example the adjective «hor»(free) becomes «hora»in feminine, «horin»in masculine plural and «horat»in feminine plural. However there are irregular adjectives that don't respect these rules partially or completely. For example the adjective «kbir»change the stem in the masuline plural and becomes «kbar».

When the noun is declined according to the four forms (masculine, feminine, feminine plural and masuline plural), then these rules apply in certain cases, like «deri»which become «deria»in feminine, «drari»in masculine plural and «deriat»in feminine plural. However generally the noun is either masculine or feminine and it accepts only the plural. For example «meftah»is a masculine noun, and it is modified only to the plural form which is «mfatah». Generally, the noun is a irregular word, because it exists various suffix of plural form and in the most cases the plural form change the stem of the noun.

2.4 Specificities syntactic dimension

The *agglutination* is the attachment of clitics (where each one has an independent syntactic function (POS)) to words in a specific order Hamdi (2015). The Algerian Arabizi is strongly agglutinated. Indeed, it can form a complete sentence, so in general a word can concatenate multiple POS tags. There are basically two types of agglutinations: verb-negation and verb-complement. The negation of the verb is formed by adding the two affixes «ma »and «ch»respectively at the beginning and the end of the conjugated verb. Thus, the negation of the verb «nakoul»(I eat) is formed as: «manakoulch»(I do not eat). So this word

is composed of several parts which belong to the POS tags: «adverb »is POS of «ma», «verb»is POS of «nakoul»(which is a conjugation in the present tense with the first person singular) and «adverb»is POS of «ch». Regarding the second agglutination type, the word includes a verb, a direct object (DO) and / or an indirect object (IO). For example, the word «ymedouhali»(they give it to me), can be segmented into three parts: «ymedou», «ha»and «li». The first part «ymedou»(which is a conjugation in the present tense with the third person plural) has a POS «verb». The second part «ha»has a POS «DO»(for the direct object pronoun which means it). The last part «li»has a POS «IO»(for the indirect object pronoun, which means to me).

3 Related works

3.1 Building resources for the Algerian Arabic

The Algerian Arabizi is an underresourced language. There are very few works in resource building. They deal either the building of lexicons or corpora. Authors in Cotterell et al. (2014) built an Algerian Arabizi-French code-switched corpus. This corpus contains discussion including domestic politics, international relations, religion, and sporting events. It is annotated for each word to distinguish French from Arabic. Abidi and Smaïli (2018) built automatically from Youtube an Algerian dialect lexicon. Each entry of this lexicon whether transcribed in Abjadia or latin, it is associated to its different transliterations. An entry may have a minimum of one transliteration and a maximum of 71. With an iterative multilingual word embeddings, they built a dictionary of 6.947 entries. GUELLIL et al. (2017) built an algerian parallel corpus in order to automatically transliterate the latin transcription to Abjadia transcription. The Algerian Arabic in the Abjadia script has been used more in the parallel corpora in order to translate it to MSA Harrat et al. (2017) Harrat et al. (2016) Harrat et al. (2014) Meftouh et al. (2015).

3.2 Automatic processing of Algerian Arabic

The Arabic rumors detection started to receive more attention in the last decade, and many detection approaches were proposed Nagoudi et al. (2020) Al-Yahya et al. (2021) Alkhair et al. (2019) Mahlous and Al-Laith (2021). However to the best of our knowledge, no work detect rumors in Algerian Arabizi. Algerian Arabizi has been processed much more for sentiment analysis. Bettiche et al. (2018) used a hybrid approach by combining the lexical based and machine learning based approaches to classify messages in positive or negative. The lexical based approach is used to automatically annotate the training corpus. In their sentiment analysis work dealing with Arabizi, Chader et al. (2019) use the transliteration and the vowels removal, to overcome phonetic and orthographic varieties. This process allows lifting the F-score of Support Vector Machine (SVM) from 76% to 87%. Mataoui et al. (2016) analysed sentiments in

Algerian Arabic with Abjadia script and latin script. They built three lexicons: keywords lexicon, negation words lexicon and intensification words lexicon, to detect text polarity. Guellil et al. (2018) built automatically a corpus containing 8000 messages written in Algerian Arabic with both Abjadia and Latin scripts using a sentiment lexicon for sentiment analysis. Abidi and Smaïli (2021) proposed a lexicon-based approach to analyse the sentiments of Maghrebi (Algerian, Moroccan and Tunisian) comments. They built a multi-script (Latin script and Abjadia script) and multilingual (Maghrebi dialects, MSA, French and English) sentiment lexicon. Firstly the lexicon is built with a small list of terms that are labelled as positive or negative manually. Secondly each word added to the lexicon is attached to the dominant polarity of the closest words in the list, in terms of distance.

Algerian Arabic has been a topic of dialect identification works. In Harrat et al. (2015) the authors have worked on the identification of several Arabic dialects: two varieties of the Algerian dialect (Algiers and Annaba), Tunisian, Syrian and Palestinian dialect. For this, they built a parallel lexicon for the five dialects and the MSA. Saadane et al. (2017) designed an identification system that supports Algerian, Moroccan, Tunisian and Egyptian. For this purpose, they adopted - similar to Harrat et al. (2015) - two approaches: a machine learning-based approach and a dictionary-based approach. The results reported in both contributions show that the dictionary-based approach performs better than the learning-based approach.

3.3 Works on morphosyntactic analysis approaches

The morphosyntactic analysis consists labelling POS and lemmatizing words. Morphosyntactic analysis systems can be grouped into five main categories: methods based on finite state automata, rule-based systems, statistical systems, hybrid systems and systems based on neural networks.

Several works dealt the morphosyntaxic analysis with different languages. Morphosyntactic analyser was built in Brill (1992) Schmid (1994) dealing with languages with latin script, in Habash et al. (2009) Gahbiche-Braham et al. (2012) Mansour (2010) dealing with MSA and in Chiang et al. (2006) Hamdi et al. (2015) to analyse Arabic dialect with Abjadia script.

However, a few works built a morphosyntaxic analyser to Arabizi. Two of them concern the Algerian Arabizi that we present in the next section. The third dealt the Tunisian Arabizi. Gugliotta and Dinarelli (2022) proposed a prediction system, based on a multi-task architecture for annotating multiple categories linguistically an Arabizi Tunisian corpus. These categories include, among others, the POS tagging and lemmas of words. The neural architecture learns to factorize information across tasks, when employing different modules with different tasks, which are learned jointly and interdependently. Each module corresponds to one decoder intended for an annotation task.

3.4 Works on morphosyntactic analysis approaches: Algerian Arabizi

Guellil and Azouaou (2016) proposed a lemmatizer and a POS tagger for Algerian Arabizi. The study, however does not deal with all the grammatical rules of Algerian Arabizi. The morphosyntactic analyser is implemented as a finite state automaton. This automaton requires a lexicon of lemmas in Arabizi and a set of linguistic rules. A term is correctly labelled if it is recognised by the automaton and arrives at a finite state. The automaton returns a word, their clitics and POS tag of each part.

Müller et al. (2020) proposed to use the multilingual language models to process Algerian Arabizi. The goal is to use transfer learning approaches to build natural language processing tools, with a low resource language. The authors show that mBert model can transfer to Algerian Arabizi in two tasks, POS tagging and dependency parsing. They turn mBERT model into a POS tagger by appending a softmax on top of its last layer.

4 rumor detection approach in Algerian Arabizi

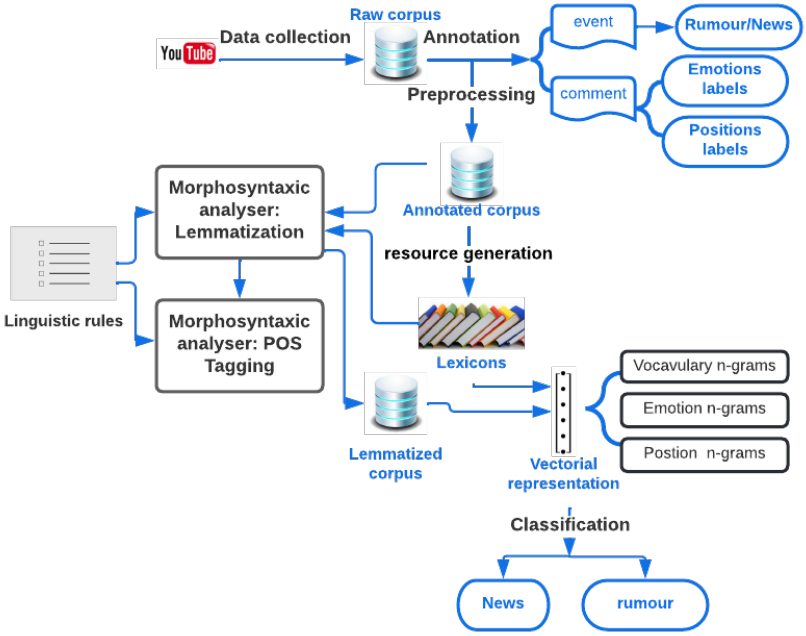

Fig. 1: rumor detection approach in Algerian dialect

Our approach tackles the issue of detecting rumors, by classifying social networks text written in Algerian Arabizi into rumor or news (see figure 1). The text corpus used is composed of comments collected from Youtube. These comments are grouped by events, where each one concerns a given topic, such as «The end of the world in 2012»which is a rumor, and «Bouteflika withdraws from the 2019 elections»which is news. Therefore each event is classified as either a rumor or a news item. rumors events are chosen from events that are already proved as fake news. The text is tagged with two different labels. One label describes the position of the author toward the topic of the event and the other shows all emotions expressed in the text, by the author. Before the classification step, a morphosyntactic analysis of the text is performed. The analysis allows lemmatizing and assigning POS tags the Algerian Arabizi. Many lexicons are built manually to represent the text for the classification process. These are the lexicons of the emotions and the positions. Several vectorial representation schemes which combining ngrams in multiple ways are built. The ngrams may consist of the words of the whole corpus, the words of the emotion lexicon and/or the words of the position lexicon. Several classification models have been built to predict rumors, namely the Naive bayes (NB), the K-Nearest Neighbors (KNN) and the SVM classifiers.

5 Arabizi morphosyntactic analyzer

Arabizi like languages, must go through linguistic processing, to improve training efficiency for classifying rumors. These processing include, among other tasks lemmatization, which allowing to reduce the inflected forms of a word, by replacing them with their lemma. For example, the lemma «kdeb»(which is the conjugation of the verb lie, in the past tense with the third person singular) replaces the following forms: «kedbou»(the verb lie conjugated in the past tense with the third person plural), «kdebtou»(the verb lie conjugated in the past tense with the second person plural), etc. This allow increasing the occurrence frequency of this lemma, and hence it helps building a better classification model.

The morphosyntactic analyzer operates in two ways, for the word and for the sentence. For the word it uses lexicons and linguistic rules to lemmatize words. For the sentence, it assigns the POS tags, when the latter cannot find the words in its lexicons.

5.1 Word based morphosyntactic analyzer

The lemmatizer identifies the inflectional form (plural, feminine, conjugation, etc.) of the word, in order to return the prefix, the lemma and the suffix. In addition, it assigns to lemma the label of the lexicon, to which it belongs. This label corresponds to POS tags: noun, verb, adjective, adverb, etc.

We modeled the linguistic rules of lemmatizer through a finite state automaton (see figure 2). We have considered the following cases:

– The verb conjugation in the four tenses (present, past, future and imperative), with the all personal pronouns.
– The negation agglutinated to verb.
– The DO and IO pronouns agglutinated to verb.
– The inflected forms terminations (feminine, masculine plural and feminine plural) of noun and adjective.
– The article agglutinated to noun and adjective.
– The possessive adjective agglutinated to noun and particle.

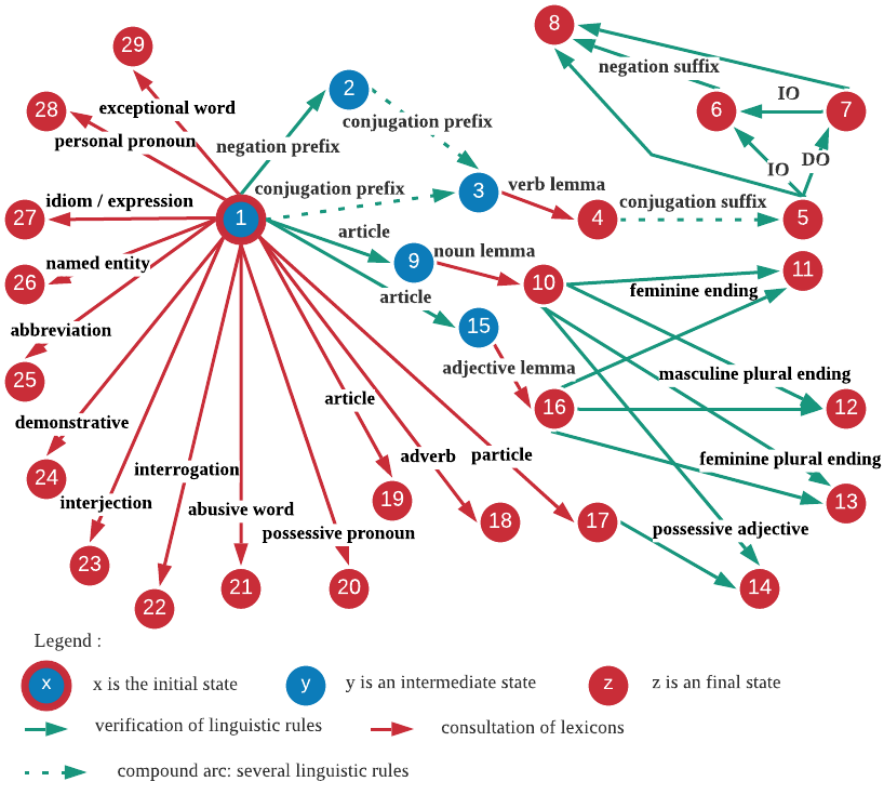

Fig. 2: Morphosyntactic analyzer for Algerian Arabizi

Automaton solves the transcription variation problem in Algerian Arabizi, through two different ways, when browsing the arcs. For the transitions that

do not require the lexicons consultation (green arcs in figure 2), we have generated all the possible transcriptions. For example, the negation suffix for the verb can have eight different transcriptions: «ch», «ech», «sh», «esh», «che», «eche», «she», «eshe». When a word analysis through this arc, we check for the presence of one of these eight transcriptions. For the transitions requiring the lexicons consultation (red arcs in figure 2), we used similarity measurement techniques between an analyzed term and a lexicon lemma, because it is tedious to generate the different transcriptions for all lemmas. We calculate the phonetic similarity between the two words using the «soundex system»Russell (1918). The words that have close phonetics will have the same soundex code, so we can represent the different transcriptions of a word with a single code. For example the different transcriptions «wektech», «waktach», «wektach», «waktech», «wektesh», «waktash», «wektash», «waktesh»(meaning when) all have the soundex code «K32».

Table 1: Lexicons

Lexicon	**Size**	**Content and/or examples**
Noun	1.035	regular and irregular nouns
Adjective	485	regular and irregular adjectives
Verb	510	verbs conjugated in the imperative and/or past tense, with the second person singular
Interrogation	55	«wach»(what), «weqtach»(when), «win»(where), «etc.
Possessive adjectives	20	«dyalek»(yours), «dyali»(mine), «dyalna»(ours), etc.
Personal pronouns	25	«ana»(I), «nta»(you), «houma»(they), etc.
Named entities	1.225	Boudiaf, Tassili, etc.
Demonstrative	25	«hada»(this one), «hadak»(that one), «lhih»(over there), etc.
Adverb	80	«lbareh»(yesterday), «lfoq»(up), «bchwiya»(slowly), etc.
Article	25	«el»(the)
Particle	145	words that make the connection between the elements of the sentence, like «bessah»(but)
Interjection	40	invariable words which express emotions like «hey», «kiw»(which means the author of comment lie)
Exceptional words	40	words relating to religion and belief like «Rebi»(God)
Idioms and expressions	160	«yetnahw gaa»(everyone will be fired)
Abbreviation	75	«Boutef»for «Bouteflika»
Abusive words	50	«bhim»(beast)

However, phonetic similarity is not enough, because several words which have different meanings can have the same soundex code, such as the code «N24»which corresponds to the words «nqol»(to say) and «nakol»(to eat). We therefore measure the orthographic similarity between the two words using the Levenshtein distance Levenshtein (1966). The closer the result is to 1, the more similar the two words are. For example the two words «nakol»and «nqol»which have the same soundex code «N24», their orthographic similarity according to the Levenshtein distance, gives 0.66, thus we deduce that these two words are different. After performing several tests, we therefore confirm that a term analyzed in the automaton is similar to a lexicon lemma, if they have the same soundex code and an orthographic similarity greater than 0.8.

Automaton search for lemmas in twelve lexicons, where each one represents a POS tag of words, in Algerian Arabizi. We manually collected nearly 4.000 lemmas (see table 1). The POS tags of the lexicons built are: noun, adjective, verb, interrogation, possessive pronouns, personal pronouns, named entities, demonstrative, adverb, exceptional words, abbreviation, article, and particle. Four other lexicons are built for regrouping idioms, interjections, exceptional words and abusive words.

5.2 Sentence based morphosyntactic analyzer

The sentence based morphosyntactic analyzer is a parser and must take over when the lemmatizer can no longer give the POS tags because the word does not exist in the lexicons. It is based on HMMs with Viterbi algorithm. For the training of the HMMs model we have manually built a corpus composed of about 600 sentences in Algerian Arabizi. These sentences have been manually labeled after lemmatization.

6 Classification and data

Before discussing the results of the experiments, we present the corpus and the different text representation approaches. We analysed our corpus to study the possible links between all the categories of annotation.

6.1 Presentation of the corpus

The corpus is organized into two groups, where one encompasses the other. The event group is composed of titles of videos that are created on Youtube, concerning a given topic, such as «La mort de Bouteflika»(The death of Bouteflika) and «La mort du président Abdelaziz Bouteflika»(The death of President Abdelaziz Bouteflika). These titles concern a single topic which is the death of Bouteflika and it is an event in the corpus. The event is annotated as rumor or as news. The comment is annotated according to the author's position to the event veracity and his emotions. The position is represented through one of the four labels:

– Support (S): when the author of the comment defends the content of the video which is annotated as a rumor or as a news
– Deny (D): when the author refutes the content of the video
– Query(Q): when the author is skeptical of the content and seeks more information
– Comment (C): when the author is neutral and expresses no reaction on the content of the video

Regarding the emotions we used the basic emotions (Anger, Fear, Joy, Sadness, Surprise, Disgust) of Ekman Ekman (1999) and we added two emotions that are expressed often by Algerians in social networks, namely Sarcasm and Schadenfreude.

The corpus is made up of 11.822 comments. It contains comments written in French, Algerian Arabizi, Algerian dialect with Abjadia script, Tamazight [7], English, MSA, and in a code-switched format. It also contains comments written only using emojis (see figure 3). The pie chart shows that Algerians use the Latin script (for the Algerian Arabizi, French and English with a total of 49.98%) more than the Abjadia script (for the MSA and the Algerian dialect with a total of 17.88%). We also notice that Algerians code-switch (about 31.32%) with the same importance as French (with 31.68%). Regarding code-switched sentences, 98% of them combine Algerian Arabizi and French.

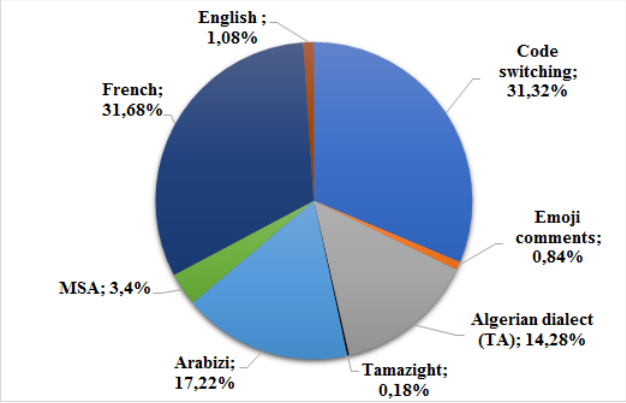

Fig. 3: Breakdown of comments by language

Before classifying the text, we have removed comments that contain the Abjadia transcription, English language, and Tamazight. Henceforth the corpus contains comments in Algerian Arabizi, French and code-switching (combining

[7] Tamazight is a language common to several countries in North Africa and the Sahel. It is carried out in different ways depending on the countries and regions

these last two languages). Consequently, the corpus has been reduced to 9.528 comments. These comments are grouped into 83 events. The events are divided between 41 rumors and 42 news. Figure 4 presents the proportion of SDQC and emotion categories in the corpus. On the two types of categories, the corpus is not balanced. For the SDQC classes, the Comment category dominates the other three. For the emotional classes, the dominant category is Anger. In several comments the authors express several emotions. However, there are about 5.299 comments where no emotion is expressed.

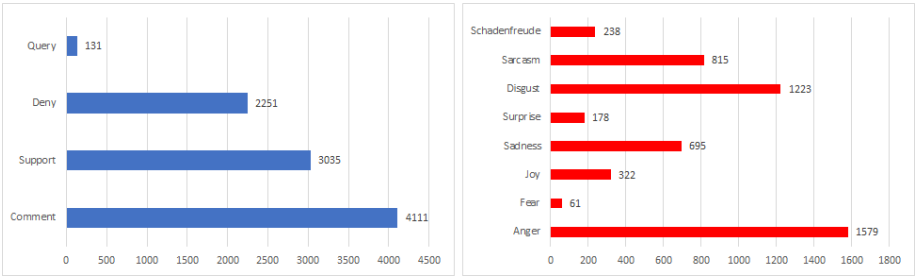

Fig. 4: Breakdown of comments according to SDQC and emotions classes

Table 2: Breakdown of comments according to emotions and SDQC categories in rumor (R) and news (N)

Emotions	**R (%)**	**N (%)**	**SDQC**	**R (%)**	**N (%)**
Anger	36	64	**Support**	22	78
Joy	30	70	**Deny**	77	23
Sadness	17	83	**Query**	55	45
Surprise	30	70	**Comment**	40	60
Disgust	37	63			
Sarcasm	36	64			
Fear	25	75			
Schadenfreude	43	57			

In order to analyze the importance of the emotions and SDQC classes for rumors and news, we studied the distribution of comments among all these classes. Table 2 presents the breakdown of comments according to emotions and SDQC classes, in rumors and news. It shows that the Support (with a percentage of 78%) and Comment (with a percentage of 60%) classes are much more expressed than the others, in the news. On the other hand in the rumors, we find much more the classes Deny (with a percentage of 77%) and Query (with

a percentage of 55%). This shows that SDQC classes can be discriminating in the classification of rumors and news. However, it can also be due to the nature of the events. For example the rumor «Facebook paid»has generated a lot of rebuttals because of the importance of the topic.

Emotions seem less convincing than SDQC classes. Indeed, table 2 shows that all emotions are more expressed in news compared to rumors. This means that their expression can depend on the nature of the topic, consequently a priori the emotions may not be descriminating in the classification of rumors and news.

6.2 Text representation schemes and classification

In order to build a good classification model and find the one that better distinguishing between rumor and news, we tested several vectorial representations using ngrams of SDQC and emotions. For building their lexicons, we manually collected 431 SDQC ngrams and 752 emotion ngrams (see table 3). Ngrams are unigrams, bigrams, trigrams and fourgrams. Each descriptor of these lexicons can be in three forms. For example the descriptor «is it true ?»is expressed in French «c'est vrai ?», in Algerian Arabizi «hadi sah»and in a code-switching form as «est ce que had lekhbar s7i7 ?». In emotion lexicon some descriptors have been duplicated. The verbs are duplicated with all the DO and IO pronouns. And the nouns are duplicated with all the possessive adjectives. This therefore extends the lexicon of emotions descriptors to 6.246 elements.

Table 3: Size of SDQC and emotion lexicons

Emotions	**Size**	**SDQC**	**Size**
Anger	286	**Support**	189
Joy	86	**Deny**	114
Sadness	106	**Query**	30
Surprise	57	**Comment**	98
Disgust	97		
Sarcasm	37		
Fear	34		
Schadenfreude	49		

We represented the text of each event with the ngrams weighted by the TFIDF. We used several combinations with ngrams of corpus vocabulary, emotion ngrams and/or SDQC ngrams. Table 4 presents the differents representation schemes used for converting the text to TFIDF value. For example representation scheme no 5 show that, first all unigram of corpus vocabulary that appear in the text are weighted by the TFIDF. Secondly all bigrams of corpus vocabulary that appear in the text are also considered. In the end the TFIDF value of all trigrams of SDQC lexicon that appear in the text are computed and concatenated with the TFIDF value of the other steps.

Table 4: Text representation schemes used

No	Representation schemes
1	Unigrams of vocabulary
2	Unigrams of vocabulary + bigrams of vocabulary
3	Unigrams of vocabulary + bigrams of vocabulary + SDQC trigrams
4	Unigrams of vocabulary + bigrams of vocabulary + emotion trigrams
5	Unigrams of vocabulary + bigrams of vocabulary + SDQC trigrams + emotion trigrams
6	Unigrams of vocabulary + bigrams of vocabulary + trigrams of vocabulary
7	Unigrams of vocabulary + bigrams of vocabulary + trigrams of vocabulary + SDQC fourgrams
8	Unigrams of vocabulary + bigrams of vocabulary + trigrams of vocabulary + emotion fourgrams

To test the contribution of the morphosyntactic analyzer, we classified the events without and with lemmatizing the text of their comments. Classification models are built using the SVM, the KNN and the NB classifiers.

7 Experimentation and results

7.1 Evaluation of the morphosyntactic analyser

We evaluated the ability of the morphosyntactic analyser to well lemmatize and assign POS tags to words. To do this, we first built a test corpus which contains 200 code-switched sentences. Secondly, we manually checked the labelling POS tags and lemmatizing done by the analyser of the words belonging to the 200 sentences. Of the 200 sentences, the morphosyntactic analyser was able to lemmatize and label 96% of the test corpus. Concerning the remaining, the morphosyntactic analyser failde in lemmatizing or labelling a single word in a sentence. For example the sentence «ya rab 3afina hna w awladna »(oh god bless us and our children), gives the lemmatization: ['ya', 'rab', 'afi', 'hna', 'w', '<unknown>'] and the POS tags: ['INTJ', 'EXCEPT', 'VERB', 'PRO', 'PART', 'NOUN']. In this example, all the terms were lemmatized an labelled by the lemmatizer except the term «awladna». It is not appearing in the lexicons. However the parser took over and correctly labeled the word, thanks to the grammatical constructions learned. This contributed to the enrichment of the noun lexicon with the word «awladna».

We also evaluated the ability of the lemmatizer to recognize words. To do this, we calculated the word recognition rate of the 9.528 comments (see results in the table 5). Table shows that more than 48.925 words (corresponding to 68.99%) in Algerian Arabizi were recognized by the lemmatizer.

Table 5: Recognition of the corpus words by the lemmatizer

Corpus data	**Lemmatizer recognition**
Total number of words	188.028
Number of French words	116.299
Number of Algerian Arabizi words	70.912
Number of Algerian Arabizi words recognized	48.925

7.2 Evaluation of classification models

Our first experiments allowed us to evaluate the contribution of the morphosyntactic analyser of Algerian Arabizi to the task of rumor classification. The table 6 presents the results of the events classification into rumors and news, with unigram vectorial representation weighted by the TDIDF. We have presented the results without and with lemmatization of Algerian Arabizi words. We note that all models gave an improvement of the F-score, after lemmatisation of the words. Indeed, the lemmatization process allowed us to increase the frequencies of Algerian Arabizi words in the corpus and thus to better build the models. In terms of model quality, KNN is ahead of the other (SVM and NB).

Table 6: Result of the events (rumor: R and News: N) classification using unigrams, without and with lemmatization

		Precision (%)		**Recall (%)**		**F (%)**	
Model	**lemmatization**	N	R	N	R	N	R
SVM	without	50	40	66.66	25	57.14	30.76
	with	77.77	75	77.77	75	77.77	75
KNN	without	62.5	55.55	55.55	62.5	58.82	58.82
	with	100	72.72	66.66	100	80	84.21
NB	without	58.33	60	77.77	37.5	66.66	46.15
	with	70	71.42	77.77	62.5	73.58	66.66

We tested several vectorial representation schemes of ngrams, using the corpus vocabulary, the lexicons of emotions and the lexicons of SDQC categories. The table 7 presents added value of the vectorial representation schemes with the SDQC and/or emotions ngrams compared to the vectorial representation with unigrams of the corpus vocabulary. These three schemes gave the best results in terms of precison, recall and F scores.

The values of F-score shows that with SVM model, all the vectorial representations allowed us to improve the classification of rumor and news. But the best one, who improve the classification with all models, is the one where we combine the unigrams and bigrams of the corpus vocabulary with the trigrams of the

SDQC trigrams. Using of the trigrams of emotions alone or with the SDQC trigrams has built a lower quality classification model. In fact, during the study of the breakdown of emotion and SDQC expressions according to rumors and news events, we predicted that the expression of SDQC makes it possible to better distinguish between rumors and news.

Table 7: Result of the events (rumor: R and News: N) classification using the best ngrams representations

		Precision (%)		**Recall (%)**		**F (%)**	
Model	**Representation**	N	R	N	R	N	R
SVM	unigrams of corpus vocabulary	77.77	75	77.77	75	77.77	75
	uni and bigrams of corpus vocabulary + Emotion trigrams	100	72.72	66.66	100	80	84.21
	uni and bigrams of corpus vocabulary + SDQC trigrams + Emotion trigrams	100	72.72	66.66	100	80	84.21
	uni and bigrams of corpus vocabulary + SDQC trigrams	100	88.88	88.88	100	94.11	94.11
KNN	unigrams of corpus vocabulary	100	72.72	66.66	100	80	84.21
	uni and bigrams of corpus vocabulary + Emotion trigrams	100	66.66	55.55	100	71.42	80
	uni and bigrams of corpus vocabulary + SDQC trigrams + Emotion trigrams	100	66.66	55.55	100	71.42	80
	uni and bigrams of corpus vocabulary + SDQC trigrams	90	100	100	87.5	94.73	93.33
NB	unigrams of corpus vocabulary	70	71.42	77.77	62.5	73.58	66.66
	uni and bigrams of corpus vocabulary + Emotion trigrams	57.14	66.66	88.88	25	69.56	36.36
	uni and bigrams of corpus vocabulary + SDQC trigrams + Emotion trigrams	57.14	66.66	88.88	25	69.56	36.36
	uni and bigrams of corpus vocabulary + SDQC trigrams	69.23	100	100	50	81.81	66.66

8 Conclusion

We presented a rumors detection approach in Algerian Arabizi. It classifies an event on a social network as rumor or news based on the positions and emotions expressed. An Algerian Arabizi lemmatizer has been proposed to improve the classification of events, but also to use it in other research contexts, to be able to linguistically process the Algerian Arabizi. It uses linguistic rules for dividing words into roots and affixes. It assigns to each word the POS tag of the lexicon in which it finds it. An Algerian Arabizi parser has been proposed, for labelling POS tags when the lemmatizer fails to do it. The lemmatizer depends on the richness of their syntaxic lexicons. When the word is not found, the parser intervenes for assigning the POS tag to word according to its position in the sentence using HMMs, and then it enriches the corresponding lexicon with this word. For the classification of the text, we used several configurations of vectorial representations by combining the ngrams of two domain lexicons: emotions and positions. Our experiments showed that the lemmatization of Algerian Arabizi considerably improved the detection of rumors. They also showed that positions are more discriminating than emotions in the classification of rumors and news.

Bibliography

Abidi, K. and Smaili, K. (2017). An empirical study of the algerian dialect of social network. In *ICNLSSP International Conference on Natural Language, Signal and Speech Processing*.

Abidi, K. and Smaïli, K. (2018). An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings. In *11th edition of the Language Resources and Evaluation Conference, LREC 2018*, Miyazaki, Japan.

Abidi, K. and Smaïli, K. (2021). Creating multi-scripts sentiment analysis lexicons for Algerian, Moroccan and Tunisian dialects. In *7th International Conference on Data Mining (DTMN 2021) Computer Science Conference Proceedings in Computer Science & Information Technology (CS & IT)*, Copenhagen, Denmark.

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic transliteration of romanized dialectal arabic. In Morante, R. and Yih, W., editors, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 30–38. ACL.

Al-Yahya, M., Al-Khalifa, H., Al-Baity, H., AlSaeed, D., and Essam, A. (2021). Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches. *Complexity*, 2021:1–10.

Alkhair, M., Meftouh, K., Othman, N., and Smaïli, K. (2019). An Arabic Corpus of Fake News: Collection, Analysis and Classification. In *Arabic Language Processing: From Theory to Practice7th International Conference, ICALP 2019, Nancy, France, October 16–17, 2019, Proceedings*, volume Communications in Computer and Information Science book series (CCIS, volume 1108), pages 292–302.

Bettiche, M., Mouffok, M. Z., and Zakaria, C. (2018). Opinion mining in social networks for algerian dialect. In Medina, J., Ojeda-Aciego, M., Galdeano, J. L. V., Perfilieva, I., Bouchon-Meunier, B., and Yager, R. R., editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications - 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part III*, volume 855 of *Communications in Computer and Information Science*, pages 629–641. Springer.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., and Rambow, O. (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 Language Processing (ANLP)*, pages 93–103, Doha, Qatar. Association for Computational Linguistics.

Brill, E. (1992). A simple rule-based part of speech tagger. In *3rd Applied Natural Language Processing Conference, ANLP 1992, Trento, Italy, March 31 - April 3, 1992*, pages 152–155. ACL.

Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

Chader, A., Lanasri, D., Hamdad, L., Belkheir, M. C. E., and Hennoune, W. (2019). Sentiment analysis for arabizi: Application to algerian dialect. In Fred, A. L. N. and Filipe, J., editors, *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2019, Volume 1: KDIR, Vienna, Austria, September 17-19, 2019*, pages 475–482. ScitePress.

Chiang, D., Diab, M. T., Habash, N., Rambow, O., and Shareef, S. (2006). Parsing arabic dialects. In McCarthy, D. and Wintner, S., editors, *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.

Cotterell, R., Renduchintala, A., Saphra, N., and Callison-Burch, C. (2014). An algerian arabic-french code-switched corpus. In *Proceedings of the First Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*. Association for Computational Linguistics.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Gahbiche-Braham, S., Bonneau-Maynard, H., Lavergne, T., and Yvon, F. (2012). Joint segmentation and POS tagging for Arabic using a CRF-based classifier. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2107–2113, Istanbul, Turkey. European Language Resources Association (ELRA).

Guellil, I., Adeel, A., Azouaou, F., and Hussain, A. (2018). Sentialg: Automated corpus annotation for algerian sentiment analysis. In Ren, J., Hussain, A., Zheng, J., Liu, C., Luo, B., Zhao, H., and Zhao, X., editors, *Advances in Brain Inspired Cognitive Systems - 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings*, volume 10989 of *Lecture Notes in Computer Science*, pages 557–567. Springer.

Guellil, I. and Azouaou, F. (2016). ASDA : Analyseur Syntaxique du Dialecte Algérien dans un but d'analyse sémantique. In *RFIA-RJCIA 2016*, volume 240.p of *Actes IA 2016*, Clermont ferrand, France. Association Française pour l'Intelligence Artificielle.

GUELLIL, I., Azouaou, F., Abbas, M., and Fatiha, S. (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*, Prague, Czech Republic.

Gugliotta, E. and Dinarelli, M. (2022). Tarc: Tunisian arabish corpus first complete release.

Guibon, G., Ermakova, L., Seffih, H., Firsov, A., and Le Noé-Bienvenu, G. (2019). Multilingual Fake News Detection with Satire. In *CICLing: International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Habash, N., Rambow, O., and Roth, R. (2009). Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*.

Hamdi, A. (2015). *Traitement automatique du dialecte tunisien à laide doutils et de ressources de larabe standard : application à létiquetage morphosyntaxique*. PhD thesis, AIX-MARSEILLE UNIVERSITÉ.

Hamdi, A., Nasr, A., Habash, N., and Gala, N. (2015). POS-tagging of Tunisian dialect using Standard Arabic resources and tools. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 59–68, Beijing, China. Association for Computational Linguistics.

Hamidian, S. and Diab, M. T. (2019). Rumor detection and classification for twitter data. *CoRR*, abs/1912.08926.

Harrat, S., Meftouh, K., Abbas, M., Hidouci, W.-K., and Smaïli, K. (2016). An Algerian dialect: Study and Resources. *International journal of advanced computer science and applications (IJACSA)*, 7(3):384–396.

Harrat, S., Meftouh, K., Abbas, M., Jamoussi, S., Saad, M., and Smaïli, K. (2015). Cross-dialectal arabic processing. In Gelbukh, A. F., editor, *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I*, volume 9041 of *Lecture Notes in Computer Science*, pages 620–632. Springer.

Harrat, S., Meftouh, K., Abbas, M., and Smaïli, K. (2014). Building resources for algerian arabic dialects. In Li, H., Meng, H. M., Ma, B., Chng, E., and Xie, L., editors, *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 2123–2127. ISCA.

Harrat, S., Meftouh, K., and Smaïli, K. (2017). Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid. In *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, Budapest, Hungary.

Joshi, A. K. (1985). *Processing of sentences with intrasentential code switching*, page 190205. Studies in Natural Language Processing. Cambridge University Press.

Khanam, Z., Alwasel, B. N., Sirafi, H., and Rashid, M. (2021). Fake news detection using machine learning approaches. *IOP Conference Series: Materials Science and Engineering*, 1099(1):012040.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Li, Q., Zhang, Q., and Si, L. (2019). Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Mahlous, A. R. and Al-Laith, A. (2021). Fake news detection in arabic tweets during the covid-19 pandemic. *International Journal of Advanced Computer Science and Applications*, 12(6).

Mansour, S. (2010). Morphtagger: Hmm-based arabic segmentation for statistical machine translation. In Federico, M., Lane, I. R., Paul, M., Yvon, F.,

and Mariani, J., editors, *2010 International Workshop on Spoken Language Translation, IWSLT 2010, Paris, France, December 2-3, 2010*, pages 321–327. ISCA.

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

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaïli, K. (2015). Machine translation experiments on PADIC: A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015*. ACL.

Müller, B., Sagot, B., and Seddah, D. (2020). Can multilingual language models transfer to an unseen dialect? A case study on north african arabizi. *CoRR*, abs/2005.00318.

Nagoudi, E. M. B., Elmadany, A. A., Abdul-Mageed, M., Alhindi, T., and Cavusoglu, H. (2020). Machine generation and detection of arabic manipulated and fake news. *CoRR*, abs/2011.03092.

Pathak, A. R., Mahajan, A., Singh, K., Patil, A., and Nair, A. (2020). Analysis of techniques for rumor detection in social media. *Procedia Computer Science*, 167:2286–2296. International Conference on Computational Intelligence and Data Science.

Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1589–1599, USA. Association for Computational Linguistics.

Russell, R. C. (1918). A method of phonetic indexing.

Saadane, H., Nouvel, D., Seffih, H., and Fluhr, C. (2017). Une approche linguistique pour la détection des dialectes arabes. In *2017-06-26*, Orléans, France.

Schmid, H. (1994). Part-of-speech tagging with neural networks. In *15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994*, pages 172–176.

Tu, K., Chen, C., Hou, C., Yuan, J., Li, J., and Yuan, X. (2021). Rumor2vec: A rumor detection framework with joint text and propagation structure representation learning. *Information Sciences*, 560:137–151.