Annotated Corpus with Negation and Speculation in Arabic Review Domain: NSAR

Ahmed Mahany[1]*, Heba Khaled[2], Nouh Sabri Elmitwally[3], Naif Aljohani[4], Said Ghoniemy[5]
Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt [1, 2, 5]
School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK[3]
Faculty of Computers and Artificial Intelligence, Cairo University, Giza 12613, Egypt[3]
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia[4]

***Abstract*—Negation and speculation detection are critical for Natural Language Processing (NLP) tasks, such as sentiment analysis, information retrieval, and machine translation. This paper presents the first Arabic corpus in the review domain annotated with negation and speculation. The Negation and Speculation Arabic Review (NSAR) corpus consists of 3K randomly selected review sentences from three well-known and benchmarked Arabic corpora. It contains reviews from different categories, including books, hotels, restaurants, and other products written in various Arabic dialects. The negation and speculation keywords have been annotated along with their linguistic scope based on the annotation guidelines reviewed by an expert linguist. The inter-annotator agreement between two independent annotators, Arabic native speakers, is measured using the Cohen's Kappa coefficients with values of 95 and 80 for negation and speculation, respectively. Furthermore, 29% of this corpus includes at least one negation instance, while only 4% of this corpus contains speculative content. Therefore, the Arabic reviews focus more on negation structures rather than speculation. This corpus will be available for the Arabic research community to handle these critical phenomena[1].**

Keywords*—Arabic NLP; negation; speculation; uncertainty; annotation; annotation guidelines; corpus; review domain; sentiment analysis

I. INTRODUCTION

Negation and speculation are commonly used linguistic phenomena, providing information on factuality and the polarity of facts [1]. Negation is a linguistic property shared by all human languages [2], which denotes the absence of something; therefore, negation affects the contextual polarity of words. On the contrary, speculative language is used to convey uncertainty about an event or idea. It means there is not enough evidence in the text to prove whether the information is 100% true. Consequently, sentences including negation or speculation may misclassify the opinionated phrases [3] or inaccurately identifying the medical terms [4], [5]. In order to efficiently identify instances of these phenomena, it is necessary to find those words expressing negation and speculation and then their scope, such as the tokens within the sentence that are affected by these cues [6]. Since negation and speculation are language-dependent, they must be addressed in all-natural languages [7]. Therefore, many studies addressed them to enhance the performance of Natural Language Processing (NLP) tasks and applications in various languages such as Sentiment Analysis (SA) [8], Machine Translation (MT) [9], and Information Extraction (IE) [5]. These studies addressed the negation and speculation scope detection using rule-based [10] and sophisticated supervised learning methods [11], [12].

Arabic Natural Language Processing (ANLP) has gained unprecedented interest in the age of big data and social media platforms, making it one of the most important research topics, especially in North Africa and the Gulf Area [13]. Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA) are the three primary forms of Arabic [14]. The Qur'an and ancient literature are written in the CA form. The MSA is mainly used in education, the official written reports like newspapers, and formal TV programs. Conversely, the DA includes all current forms of Arabic spoken, written on social media platforms, and reviewed applications and websites where it varies nationally and internationally depending on location [15]. Since the DA has no syntactic rules and multiple forms of the same word, ANLP tasks are challenging.

Negation frequently occurs in the Arabic language and is one of the dominant linguistic methods for changing the text polarity, so negation detection is highly considered in the Arabic Sentiment Analysis (ASA) [3]. However, the presence of negation words in a sentence does not imply that all the sentimental words are inverted. Still, there are odd cases where the presence of negation terms may confirm the polarity of the following lexeme [16]. In the implicit form of negation, a sentence can be negated without using negation words. The level of speculative content increases or decreases the certainty of polarity classification [17]. Few Arabic studies have addressed the impact of negation and speculation using simple rules. Hamouda and El-Taher considered the frequency of negation terms in the ASA task as a classification feature, but the effect of the negation feature on the sentiment classification was not clearly mentioned [18]. In 2015, Duwairi and Alshboul defined six handcrafted rules to handle negation in the Modern Standard Arabic (MSA) texts in the review domain to enhance the performance of the ASA [19]. Even though they addressed the MSA, which follows well-defined rules, the simplistic approach has proven inadequate for a syntactically and morphologically rich language like Arabic. El-Naggar et al. considered several valences to build a negation-aware classifier for ASA in MSA and the Egyptian dialect [20]. Later, Assiri et al. formulated four rules to handle negation in the Saudi dialect

*Corresponding Author

[1] https://github.com/amahany/NSAR

[21]. In addition, Kaddoura et al. have proposed a system that inverts the polarity of a sentence's clause if a negation term precedes a positive or negative pattern [3]. Regardless of the improvement in performance in these systems' experimental results [3], [20], [21], none handled the implicit form of negation frequently used in Arabic. Simple rule-based algorithms cannot handle all the negation and speculation cases for the various Arabic language forms and dialects [14]. According to the findings of our earlier work, the treatment of negation scope detection utilizing supervised based learning is promising [12]. To the best of our knowledge, there are no available Arabic corpora annotated with negation or speculation in various domains including the review, newswire and medical domains. Furthermore, speculation detection in ASA has not been studied in any research work.

In the last decade, there has been a growing interest in detecting negation and speculation. Nevertheless, the available open-access corpora for low-resource languages, such as the Arabic language [22], are limited compared to the English and the Spanish languages [7]. Speculation corpora are even more scarce than those for negation, with the majority focusing on the biomedical domain. Since negation and speculation are language-dependent phenomena, the negation- and speculation-aware models from other languages, such as English, cannot be applied to the Arabic text because the syntactic structure of negation in Arabic differs from that in English. Therefore, developing an annotated corpus with negation and speculation for the Arabic review domain is required. It is very important to know that negation- and speculation-aware systems improve the overall systems performance [9], [11].

The rest of the paper is organized as follows: Section II shows the different sources for our corpus. Section III details the annotation guidelines we build for the negation and speculation texts in the Arabic review domain. The annotation process and its result including the agreement analysis of the annotators and the discussion are presented in IV and V. Finally, Section VI concludes the paper and suggests the future work.

II. Corpus Collection

This section demonstrates the overall characteristics of the Negation and Speculation Arabic Review (NSAR) corpus, as well as a brief description of the texts that compromise it. Furthermore, general statistics are presented regarding each source's size and polarity distribution. The NSAR corpus is comprised of texts extracted from three well-established and benchmarked Arabic review corpora: Large Scale Arabic Book Review (LABR) [23], Large Arabic Multi-domain Resources (LAMR) [24], and Multi-domain Arabic Sentiment Corpus (MASC) [25]. Table I shows the distribution of randomly selected positive and negative sentences from each source, with 2,312 positive reviews accounting for approximately 77% of our corpus. Each topic has a different number of sentences, but the average number of words per sentence is nearly the same. The LABR corpus contains 63K book reviews, with ratings ranging from 1 to 5 stars [23]. Aly and Atiya considered the reviews with 4 or 5 stars with positive polarity and those with 1 or 2 stars with negative polarity. The authors collected these reviews from the best Arabic books listed in the social network for book readers [2]; hence, most of the randomly selected reviews are positive reviews, as per Table I. The LAMR corpus is the second source for NSAR corpus, and it consists of 33K reviews scrapped via Scrapy framework [3] from various reviewing websites, Souq [4], TripAdvisor [5], Elcinema [6], and Qaym [7] , including reviews for various items and services [24]. Each sentence includes the review text and normalized rating that could be positive, negative, or mixed polarity. The third source, MASC [25], includes 8,860 reviews on different topics such as shopping, restaurants, and software applications written in multiple Arabic dialects. These reviews were obtained primarily from Jeera [8], Qaym, Google Play, Twitter, and Facebook. The majority of the reviews from LAMR and MASC were composed in Egyptian, and Gulf areas' dialects. On the contrary, most of the LABR samples were written in the MSA form. The review texts in the NSAR corpus are collected from various sources to ensure that it captures the diversity of dialectical language usage in the review domain.

TABLE I. Corpus Statistics

Corpus	Topic	Positive	Negative	Total
LABR	Books	879 (84.52%)	161 (15.48%)	1,040
LAMR	Touristic Attractions	102 (94.44%)	6 (5.56%)	108
	Hotels	74 (100%)	0 (0%)	74
	Products	684 (75.58%)	221 (24.42%)	905
	Resturants 1	248 (74.47%)	85 (25.53%)	333
	Restaurants 2	114 (81.43%)	26 (18.57%)	140
MASC	Software	210 (52.50%)	190 (47.50%)	400
	Products	1 (9.09%)	10 (90.91%)	11
Total		**2,312 (76.79%)**	**699 (23.21%)**	**3,011**

III. Annotation Guidelines

Negation and speculation phenomena are interrelated and have similar characteristics: they both have a scope, so they affect the part of the text denoted by the presence of negating or speculative keywords. Furthermore, both of them have two types: implicit and explicit. In the case of the explicit type, the phenomenon cue is written in the sentence, whereas being understood in the case of the implicit one without a cue. Sentences including a negation cue are not necessarily annotated for negation; however, they may have speculative content. Therefore, the annotators should read sentences containing negation cues carefully. In most cases, the keywords influence their scope, aligned from the left to the end of the clause or the sentence.

The following subsections list the general principles, negation, and annotation guidelines. Furthermore, the special

[2] https://www.goodreads.com/
[3] https://scrapy.org/
[4] https://souq.com/
[5] https://www.tripadvisor.com/
[6] https://elcinema.com/
[7] https://www.qaym.com/
[8] https://www.jeeran.com/

or complex cases for both phenomena are demonstrated. In order to illustrate examples in the annotation guidelines, the negating cues are surrounded by a negation symbol (¬), the speculative cues are surrounded by an uncertainty symbol (∓), and their scope boundaries are surrounded by parenthesis.

A. General

When annotating the negation and speculation, several general rules must be followed, which are adapted from the BioScope annotation guidelines [6], then modified to the Arabic language and review domain. Sentences with some instance of negative or speculative language will be only considered. In addition, the min-max strategy should be followed during the annotation. The minimal unit (single word) that expresses the negation or speculation will be marked as a cue. Nevertheless, in some cases, a cue may include more than a single word which is called a complex cue. The maximum number of words affected by a cue will be marked as the scope for negation or speculation. The scope usually starts after the keyword and ends at the end of the phrase, clause, or sentence. However, the scope may include a word or a statement preceding it. The below list summarizes the general rules for both negation and speculation:

- A sentence may contain more than one cue instead of only one keyword; in this situation, each cue should be annotated separately.
- Structures of negation and speculation can be annotated in a single sentence.
- The cue is not included in the scope, but it may be included in complex cases and in the scope that includes words preceding and following a cue.
- If a sentence contains a cue that appears at the end of the sentence, the phenomenon's scope is limited to the cue.
- Due to the improper use of spaces in the informal Arabic text, a cue+verb/noun may be concatenated without a space; in this case, the verb/noun will be included in the negation/speculation cue.
- The coordinating conjunctions و (and) extend the scope.
- Annotators will only annotate the cue and leave the scope for the linguist expert if the annotator is unsure about the scope.
- There is an annotation element called the ‘undecided’ used if the annotator is unsure what type the keyword should be assigned.

Additionally, each type of a negation or speculation structure is depicted with an example where the transliteration and English translation of these examples are listed in Appendix I.

B. Negation Structures

- لا (*no*) is the most used negating Arabic word, which is used to deny the occurrence of a verb in the past and present tenses, as well as to deny a nominal sentence. Therefore, the scope begins with the negative cue and ends at the end of the sentence.

1) أحلام اجمل الايام كانت هناك و (ذكريات ¬ لا ¬ تنسى) مكان رائع يجمع الكل فى حركه و ضحك وحياه

- ما (*did not – does not*) interfere with past or present verbs in dialectal Arabic and the formal Arabic forms in the nominal sentences. It is most often found in a pre-verbal state. To ensure that ما with a verb in the past is working, replace it with لم (*does not*) followed by the same verb in the present.

2) لو كان سعره اقل وفيه فلاش كنت قيمته اكار من كدا فشل واضح من سوني انها ¬ ماحطتش ¬ (ليه فلاش)

- لم keyword is used with a verb in the present tense to deny a truth that occurred in the past. In exceptional cases, it may affect something in the present or future.

3) جهاز جبار ويتفوق علي نظرائة من الايباد والسامسونج بجد رابع بس ¬ لم ¬ (ياخذ حجم الدعايه المطلوبه)

- لما (*Lamma*) is used before a verb in the present to deny something in the past and may deny something in the future. However, if it comes with a verb in the past, it does not deny anything after it.
- لات (*no*) cue is used only in the classical Arabic form.
- إن (*Inn*) affects nouns, past, or present verbs. إن will be effective if it gets replaced with another cue and reverses the polarity. There is a distinction between إن و أن (*Ann*) where أن is not a negating cue. Furthermore, إن و إن may be written ان without همزة (*Hamza*) in the dialectical Arabic. Therefore, it is necessary to read the sentences carefully to determine the correct form in accordance with the context of the sentence.
- لن (*will not*) is used with a verb in the present tense to deny something in the future.

4) 7 روايات في رواية ، تهكم وسخرية وأدب وأشياء اخرى ¬ لن ¬ (تمل منها)

- ليس (*Not*) is a verb in the past tense, extending the negation to the end of the sentence. The origin of ليس is لا+ايس which means no + existing. It has different forms like لست – ليست – ليسا – لسنا – لسن – فلست

5) يستاهل لانه دعم اللغة العربية لكن ياريت يثبت البرنامج على رقم المستخدم ¬ وليس ¬ (على حساب)

- عدم (*None*), عدا (*Except*), and دون (*Without*) keywords are used to deny a noun or nominal sentence.

6) الذي فاجئني وزعجني هو ¬ عدم ¬ (وجود ريموت فيها) بالمقارنة مع السعر

- مش (*Mesh*) keyword and the pattern of م + verb + ش are mainly used in the Egyptian dialect to deny a verb.

7) تطبيق رائع وفيه خصوصيه ¬ محدش ¬ (يعرف دخلت امتي خرجت امتي) انت حتي لو بتكتب (اللي قدامك ¬ مش ¬ بيعرف)

- مفيش (*Not*) used in the Egyptian dialect to deny nouns

8) فكرني بفلاش و سماش والحاجات دي بس ¬ مفيش ¬ (كلمات متقاطعة
(

- مو (*Not*) keyword used in the Gulf/Levant countries to deny a verb, noun, or an event.

9) مرررررره حلو وعصيراته فرش ⌐ ومو ⌐ (حاطين له لاسكر ولا مويه كله فرش)

C. Speculation Structures

- There are many Arabic adjectives or adverbs that indicate speculation such as, محتمل – احتمال – ممكن – مرجح – شكاك – أحيانا

10) نصيحتي ان (في محلات فطاير تانية في حدايق حلوان ∓ ممكن ∓ تكون أفضل كتير)

- Some other adjectives, if get preceded by a negation cue, it indicates a speculative content, such as غير مؤكد

11) اول روايه اقرأها لباولو كويلهو ∓ ولااعتقد ∓ (انها الاخيره)

- This list of verbs, but not limited, indicates speculation - يعتقد - يظن - يمكن – توحى – يبدو – تظهر – تشير – يفترض – يشك. In addition to, the noun form for some of these verbs like الافتراض – فرضية – الظن – الشك

12) ⌐ لم ⌐ (استسيغ نزار فى شعر الفصحى) ∓ اعتقدوا ∓ فى الشعر الحر اكثر ابداعا)

- These Arabic particles لو – قد – ربما – لعل –ما بين indicate speculative content.

13) فندق مريح صراحة اسعارة جيدة ∓ ما بين ∓ (150 الى 300 ريال الليلة) ونظيف جدا

- Conjunction keywords such as أو (*or*) have the scope of elements ranging from the right to the left side of the conjunction. However, in instances where the conjunction is composed of two or more words like أو, إما (*Or*), سواء (*Whether*), the scope does not change.

14) أنصح بيه أي حد بيدرس (هندسة كهربية ∓ أو ∓ عايز يدرسها) هتوفر عليه جنون كتير

- Sentences starting with a question that should be annotated as speculative.

- If the speculation cue is present at the start of the sentence, then the scope extends to include the whole sentence.

15) ∓ربما ∓ (يكون الكتاب جيداً ولكن بروز شخصية الكاتب المتملقه تفسد ذلك؟)

D. Negation Complex Cases

The presence of a negation keyword does not automatically negate a sentence as follows:

- For example, إن that assures something.

16) ⌐ ما ⌐ إن رأيت ولا سمعت بمثله

- Example for ما used for wonderment.

17) ⌐ ما ⌐ هذه الرومانسية الحالمة ⌐ وما ⌐ هذا الاسلوب الناعم الجميل هذه الرواية من اجمل ⌐ ما ⌐ قرات على الاطلاق

- Question marked withليس reverts the sentence from being negated to being proven

18) ⌐ أليس ⌐ هذا بالحق

- If a sentence consists of ما then إلا, the negation is canceled

19) ⌐ مَا ⌐ هَٰذَا بَشَرًا إِنْ هَٰذَا إِلَّا مَلَكٌ كَرِيمٌ

- Two consecutive negating cues like ما cancels the effect of negation.

- غير sometimes used in the Gulf countries to express something unique.

20) جدة ⌐ غير ⌐

- غير in some cases means change but not a negation cue

21) كتاب ⌐ غير ⌐ مجرى تفكيرى خلاه اوسع خلانى أثق اوووى فى العلامات

- مش بس is used to assure something

22) تجنننننن انا بصراحه ⌐ مش ⌐ بس الفندق جميل كل حاجه زورتها كانت جميله قوي قوي قوي قوي

- Verb in forms of ما + أفعل

23) أجمل ⌐ ما ⌐ فيه هو إفطاره

In some other cases, the negation is implied in the sentence without any negating cue while understood from the context of the text.

- The sentence implies denial without any negative cues such as

24) كتاب خفيف و واقعي و ⌐ بعيد ⌐ (عن المبالغة تماما) كل شيء فيه حقيقي

- Negating the not obvious

25) (وَتَرَى النَّاسَ سُكَارَىٰ وَمَا هُم بِسُكَارَىٰ)

E. Speculation Complex Cases

Certain speculation cases are marked using few keywords.

- For example, قد can express speculative content only if the verb following it is in the present tense form, as in point 25. However, the content in point 26 is not speculative and comes with a verb in the past tense.

26) الكلمة ∓ قد∓ (تفعل فى الانسان ما لم تفعلة الادوية القوية) لك كل قدير

27) السلعة ليست بالجودة المطلوبة ∓ وقد ∓ اشتخدمتها لمرة واحدة فقط ولم ارجع لاستخدامها مرة ثانية

- When used at the end of a sentence, the negation cue ما indicates speculation.

28) جلسات رائعة جلسات المطعم الخارجية رائعة خصوصا في فصل الربيع والشتاء اما (الاكل فجيد) ∓ نوعا ما ∓

- Most of the cases that use لعل do not express speculation; however, it represents hopefulness.

29) وَجَعَلَ لَكُمُ السَّمْعَ وَالْأَبْصَارَ وَالْأَفْئِدَةَ ۙ لَعَلَّكُمْ تَشْكُرُونَ

- In some cases, speculation cues may be used to imply an affirmation.

30) الَّذِينَ يَظُنُّونَ أَنَّهُم مُّلَاقُو رَبِّهِمْ وَأَنَّهُمْ إِلَيْهِ رَاجِعُونَ

IV. NSAR Annotation

This section describes the procedure followed in the annotation process of the NSAR corpus. Initially, the guidelines are created based on the negation rules of the formal Arabic language in addition to the commonly used slang negating cues in the Egyptian and Gulf countries' dialects. Then, a list of Arabic keywords for the speculation is built which would indicate speculative content, and subsequently, these rules are applied to annotate a sample of the corpus and extract any additional cases from the corpus to enhance these rules for the annotation process.

There is a need for a tool for the annotation process to build and develop NSAR corpus. There are many available annotation tools for this purpose. Based on an evaluation of the well-known annotation tools in this study [26], WebAnno[9] is selected, which achieved the highest score [27]. WebAnno is an open-source web-based annotation tool that provides full functionality for both semantic and syntactic annotations. Furthermore, it supports adding user-defined annotation layers as we did for the negation and speculation. The user-defined layers are only supported in TSV3 format, where there is an open-source Python library to extract the annotations written in TSV[10]. As in Section II, NSAR corpus is collected from three different Arabic corpora from the review domain labeled as positive or negative and written in CSV file format. Therefore, we transformed the input files from CSV to TSV file format. Five user-defined labels associated with the WebAnno project: sentiment, negation, speculation, bad, and undecided are created. The sentiment has one feature called ‘polarity’ with ‘negative’ or ‘positive’ values, used with the transformation from CSV to TSV for the sentiment labeling. For the negation and speculation labels, every label has a tag set with two different values ‘cue’ and ‘scope’ which are associated to each other using two user-defined relations ‘NegRel’ and ‘SpecRel’. The other two labels ‘bad’ and ‘undecided’ are used to highlight any inappropriate or hateful content in the text or the annotator cannot take a decision about a sentence.

The annotation process was implemented in three phases: the first phase was to describe the annotation guidelines and train the annotators on using WebAnno, then the annotators carried out the annotation to measure the inter-annotator agreement (IAA), and finally, a linguist expert resolved the disagreements between them. Two independent Arabic native speakers carried out this process; one is an experienced annotator with a solid background, and the second is a well-trained person. Each file has been annotated by both annotators.

V. Results and Discussion

In this section, we explore the result of the annotation process. The Cohen’s Kappa coefficient [28] is used to measure the quality of the annotation process. Cohen’s Kappa of value 0.95 for the negation and 0.8 for speculation are obtained. These values demonstrate that the speculation annotation is more complex than the negation in Arabic. Table II shows the NSAR corpus, which includes 862 negated sentences out of 3,011, and only 121 sentences containing at least one speculative content.

The disagreements between the two annotators were revised by a linguist expert [6]. The majority of disagreement cases in negation are caused by common human errors, such as one of the annotators forgetting to relate the negation cue to its scope using the relation layer. Since a single sentence may contain multiple negation structures [29], this layer is added and should be specified for each annotation. The speculation cases, on the contrary, are ambiguous and may lead the annotator to consider it a negation or speculation [7]. Therefore, it had a higher level of disagreement than the negation. These cases involve an issue within the scope of speculation, such as the non-inclusion of a word. In addition to the undecided label, the disagreements have been curated by the first author and the linguist expert.

Table II shows that 29% and 4% of total sentences have at least negation and speculation structures, respectively; however, these percentages vary from topic to topic. For instance, MASC sub-corpus includes high rates of negating and speculative content.

TABLE II. NSAR Statistics

Corpus	Topic	Size	Negation	Speculation
LABR	Books	1,040	248 (23.85%)	46 (4.42%)
LAMR	Touristic Attractions	108	20 (18.52%)	1 (0.93%)
	Hotels	74	7 (9.46%)	2 (2.70%)
	Products	905	284 (31.38%)	30 (3.31%)
	Resturants 1	333	98 (29.43%)	10 (3%)
	Restaurants 2	140	33 (23.57%)	3 (2.14%)
MASC	Software	400	166 (41.50%)	28 (7.00%)
	Products	11	6 (54.55%)	1 (9.09%)
Total		**3,011**	**862 (28.63%)**	**121 (4.02%)**

The subject types in Arabic sentences change the form of most Arabic words, such as verbs ذهب (He went) and ذهبت (She went). There are other various forms of negation in Arabic that have the same meaning in English. This example shows the negation difference between the MSA and Egyptian dialect where ملكشى in the Egyptian dialect is derived from لا شىء لك or ليس لك شىء in MSA form, where all of them means (you do not own anything). Another example, مكنتش in the Egyptian dialect, which is derived from لم تكن or لم أكن in MSA, means (I do not + verb) or (She does not + verb) according to the context. However, removing a single character from this word as مكنش will change the meaning to be (He does not + verb). These examples demonstrate the complexity of negation in Arabic, especially in the dialect Arabic. Furthermore, the spelling rules are not followed in dialectical Arabic, resulting in tokenization issues such as in الكتابةلاتظهر (The written text does not appear) [3]. There is no space between the three words that should formally be used. Other instances in the dialect of Arabic include different forms for the same Arabic word with the same meaning as in مافيش and مفيش (None-existence). Therefore, we normalized the commonly used negation and

[9] https://webanno.github.io/webanno/

[10] https://github.com/neuged/webanno_tsv

speculation cues, as depicted in Table III and Table IV. The Negator لا and speculative cue لو account for approximately 45% of the negation and speculation cues, respectively.

TABLE III. THE COMMON NEGATING CUES IN NSAR

Normalized Negation Cues	**Frequency**
لا	455
ما	161
لم	129
غير	84
مش	78
لن	20
دون	25
مو	30
ليس	74
عدم	21
عدا	7
مفيش	5
ملهاش - معجبتنيش - الا	3
محدش - مقدرتش	2
مبتستاهل	2
مكنتش - مكنش - ماكنتش - معجبنيش - معجبتنيش - مفهمتش - مفهمتهاش - مبقتش - مابقتش - معرفش - معرفتش - ميستحقش - ميخلكش - ملوش - مفيهوش - ماينفعش - متنبعتش - معتش - مبقتش - مبيضحكش - منبسطش - محستهاش - متعميلهاش - بلاش - عديم - معاد - مب	1

TABLE IV. THE COMMON SPECULATION CUES IN NSAR

Normalized Speculation Cues	**Frequency**
لو	41
اعتقد	11
كانت	8
او	6
قد	6
اظن	5
ممكن	5
ربما	4
يمكن - لااعتقد - معظم - احيانا	3
اتمنى - ولا - تقريبا - لاادرى - إذا	2
مما يثيرالشك - بالرغم من الشكوك - فعلا كان - نوعا ما - لا اظن - ما بين - غالبا - ياريت - تبدو - تاكد - ان - لما	1

Table V displays the average, minimum, and maximum scope lengths for both negation and speculation for each topic. For the negation scope, the minimum and average scope lengths are nearly identical, but there is a notable variation in the maximum scope length for each topic. This notice in books and software topics usually negate the longest part of the sentence. Table V also shows that the speculated words within a sentence are longer than the negated words because the speculation structures usually affect the whole sentence, as described in the annotation guidelines.

TABLE V. NSAR NEGATION AND SPECULATION SCOPE LENGTH

Corpus	**Topic**	**Negation Scope**			**Speculation Scope**		
		Max	***Min***	***Avg***	***Max***	***Min***	***Avg***
LABR	Books	66	2	22	82	2	32
LAMR	Touristic Attractions	45	3	21	13	13	13
	Hotels	33	7	17	23	21	22
	Products	56	3	22	65	12	35
	Resturants 1	50	2	17	86	10	35
	Restaurant 2	44	3	26	48	10	27
MASC	Software	60	2	20	62	8	32
	Products	40	5	30	52	52	52
All		**66**	**2**	**21**	**86**	**2**	**31**

Table VI presents the distribution of negated and speculated sentences based on the overall polarity of the sentence. On average, the number of sentences with negation structures and positive polarity is the same as negative polarity. Nonetheless, the number of negation cases in the software topic with negative polarity is more than the cases with positive polarity. In addition, the speculative contents within positive polarity account for 66% of the corpus speculation cases as it is the majority in the books and software topics. According to our observation, the book's topic includes most negation and speculation cases, which are typically used to cancel something negative about the books. Furthermore, most of the software advantages or features are negated or speculated.

Fig. 1 and Fig. 2 demonstrate the number of negation cases in each sentence within the three sub-corpora. The number of negated sentences that include more than two negation scopes in one sentence is 173, accounting for 20% of the negation cases in the NSAR corpus. However, there are only three sentences with two speculation scopes. This finding further proves that the speculative content in the review domain includes the entire sentence as long as the polarity.

TABLE VI. NEGATION AND SPECULATION SENTENCES PER POLARITY

Corpus	**Topic**	**Negation**		**Speculation**	
		Pos	***Neg***	***Pos***	***Neg***
LABR	Books	165	83	29	17
LAMR	Touristic Attractions	17	3	1	0
	Hotels	7	0	2	0
	Products	114	170	16	14
	Resturants 1	59	39	7	3
	Restaurants 2	13	20	2	1
MASC	Software	54	112	23	5
	Products	1	5	0	1
Total		**430**	**432**	**80**	**41**

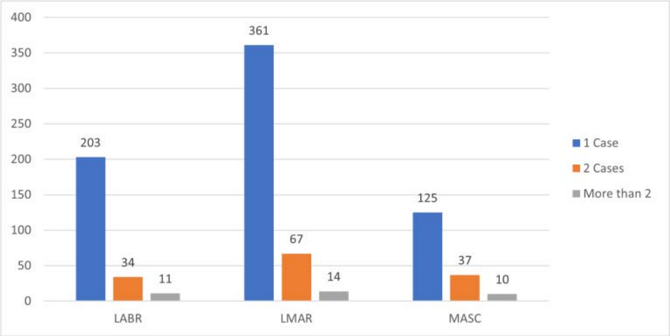

Fig. 1. The Distribution of Negation Structures per Sentence.

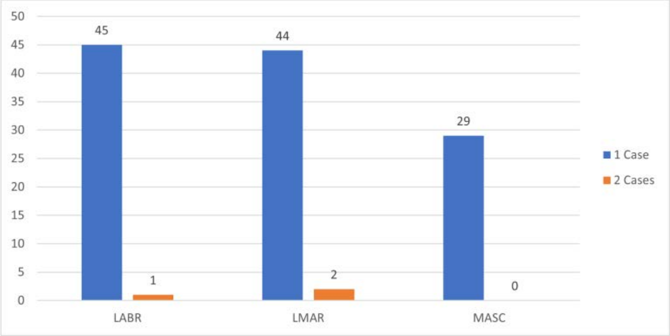

Fig. 2. The Distribution of Speculation Structures per Sentence.

VI. Conclusion and Future Work

The DA texts are used in people’s day-to-day conversations on social media platforms and review websites. Many research groups worked on the sentiment analysis task, and some of them considered the negation linguistic feature and highlighted its significance using simple rules. However, researchers still have challenges in addressing various structures of the negation phenomenon as long as the speculation. This paper presented the first Arabic corpus in the review domain annotated with negation and speculation (NSAR) to tackle these challenges using supervised learning techniques. This corpus was annotated by two Arabic native speakers who adhered to strict annotation guidelines that were reviewed by a linguist expert. The Cohen's Kappa coefficients were used to measure annotator agreement and obtained 95 and 80 for negation and speculation, respectively. The results show that the annotation guidelines were written clearly. NSAR will be made available, which will contribute to the detection of negation and speculation, as well as the sentiment analysis task. The future work includes extending the corpus by annotating the events element as long as the negation focus. In addition, we plan to apply the recent deep learning techniques on this corpus to study the impact of negation and speculation on various ANLP tasks.

References

[1] Velldal, L. Øvrelid, J. Read, and S. Oepen, “Speculation and Negation: Rules, Rankers, and the Role of Syntax,” Computational Linguistics, vol. 38, no. 2, pp. 369–410, Jun. 2012.

[2] J. H. Greenberg, “Universals of human language,” Stanford University Press, vol. 4, 1978.

[3] S. Kaddoura, M. Itani, and C. Roast, “Analyzing the Effect of Negation in Sentiment Polarity of Facebook Dialectal Arabic Text,” Appl. Sci., vol. 11, no. 11, p. 4768, May 2021.

[4] O. Solarte Pabón, M. Torrente, M. Provencio, A. Rodríguez-Gonzalez, and E. Menasalvas, “Integrating Speculation Detection and Deep Learning to Extract Lung Cancer Diagnosis from Clinical Notes,” Appl. Sci., vol. 11, no. 2, p. 865, Jan. 2021.

[5] C. Dalloux et al., “Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora,” Natural Language Engineering, vol. 27, no. 2, pp. 181–201, Mar. 2021.

[6] V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik, “The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes.,” BMC bioinformatics, vol. 9, no. 11, p. Suppl 11-S9, Nov. 2008.

[7] A. Mahany, H. Khaled, N. S. Elmitwally, N. Aljohani, and S. Ghoniemy, “Negation and Speculation in NLP: A Survey, Corpora, Methods, and Applications,” Applied Sciences, vol. 12, no. 10, p. 5209, May 2022.

[8] S. M. Jiménez-Zafra, R. Morante, E. Blanco, M. T. Martín-Valdivia, and L. Alfonso Ureña-López, “Detecting negation cues and scopes in Spanish,” in LREC 2020 - 12th International Conference on Language Resources and Evaluation, 2020, pp. 6902–6911.

[9] M. M. Hossain, A. Anastasopoulos, E. Blanco, and A. Palmer, “It’s not a Non-Issue: Negation as a Source of Error in Machine Translation,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 3869–3885.

[10] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, “A simple algorithm for identifying negated findings and diseases in discharge summaries,” Journal of Biomedical Informatics, vol. 34, no. 5, pp. 301–310, Oct. 2001.

[11] H. Fei, Y. Ren, and D. Ji, “Negation and speculation scope detection using recursive neural conditional random fields,” Neurocomputing, vol. 374, pp. 22–29, Jan. 2020.

[12] A. Mahany et al., “Supervised Learning for Negation Scope Detection in Arabic Texts,” in Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), 2021, pp. 177–182.

[13] A. Alsayat and N. Elmitwally, “A comprehensive study for Arabic Sentiment Analysis (Challenges and Applications),” Egyptian Informatics Journal, vol. 21, no. 1, pp. 7–12, Mar. 2020.

[14] N. Y. Habash, Introduction to Arabic natural language processing, 1st ed., vol. 3, no. 1. Morgan and Claypool Publishers, 2010.

[15] A. Elnagar, S. M. Yagi, A. B. Nassif, I. Shahin, and S. A. Salloum, “Systematic Literature Review of Dialectal Arabic: Identification and Detection,” IEEE Access, vol. 9, pp. 31010–31042, 2021.

[16] S. R. El-Beltagy, “NileULex: A phrase and word level sentiment lexicon for Egyptian and modern standard Arabic,” Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, pp. 2900–2905, 2016.

[17] N. P. Cruz, M. Taboada, and R. Mitkov, “A machine-learning approach to negation and speculation detection for sentiment analysis,” Journal of the Association for Information Science and Technology, vol. 67, no. 9, pp. 2118–2136, Sep. 2016.

[18] A. E.-D. Hamouda and F. E. El-taher, “Sentiment Analyzer for Arabic Comments System,” International Journal of Advanced Computer Science and Applications, vol. 4, no. 3, pp. 99–103, 2013.

[19] R. M. Duwairi and M. A. Alshboul, “Negation-Aware Framework for Sentiment Analysis in Arabic Reviews,” in 2015 3rd International Conference on Future Internet of Things and Cloud, 2015, pp. 731–735.

[20] N. El-Naggar, Y. El-Sonbaty, and M. A. El-Nasr, “Sentiment analysis of modern standard Arabic and Egyptian dialectal Arabic tweets,” in 2017 Computing Conference, 2017, pp. 880–887.

[21] A. Assiri, A. Emam, and H. Al-Dossari, “Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis,” Journal of Information Science, vol. 44, no. 2, pp. 184–202, Jan. 2018.

[22] N. Alalyani and S. Larabi, “NADA: New Arabic Dataset for Text Classification,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, pp. 206–212, 2018.

[23] M. Aly and A. Atiya, “LABR: A large scale arabic book reviews dataset,” in ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2013.

[24] H. ElSahar and S. R. El-Beltagy, “Building Large Arabic Multi-domain Resources for Sentiment Analysis,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9042, pp. 23–34, 2015.

[25] T. Al-Moslmi, M. Albared, A. Al-Shabi, N. Omar, and S. Abdullah, “Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis,” Journal of Information Science, vol. 44, no. 3, pp. 345–362, Jun. 2018.

[26] M. Neves and J. Ševa, “An extensive review of tools for manual annotation of documents,” Briefings in Bioinformatics, vol. 22, no. 1, pp. 146–163, Jan. 2021.

[27] R. Eckart de Castilho et al., “A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures,” in workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), 2016, pp. 76–84.

[28] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[29] N. Pröllochs, S. Feuerriegel, B. Lutz, and D. Neumann, “Negation scope detection for sentiment analysis: A reinforcement learning framework for replicating human interpretations,” Information Sciences, vol. 536, pp. 205–221, Oct. 2020.

APPENDIX I

TABLE VII. ANNOTATION GUIDELINES EXAMPLES

No.	Arabic Text	Transliteration*	English Translation
1	أحلام اجمل الايام كانت هناك و ذكريات لا تنسى مكان رائع يجمع الكل فى حركه و ضحك وحياه	Aḥlām ajmal al-Ayyām kānat hunāk wa Dhikrayāt lā tunsá makān rā’i‘ yajma‘u al-Kull fī ḥrkh wa ḍaḥika wḥyāh	The most beautiful days were spent there with unforgettable memories; it is a wonderful place that brings everyone together in liveliness, laughter and life.
2	لو كان سعره اقل وفيه فلاش كنت قيمته اكار من كدا فشل واضح من سوني انها ماحطتش ليه فلاش	Law kāna si‘ruhu aqall wa-fīhi Flāsh Kunt qymth akār min kdā fashal Wāḍiḥ min Sūnī annahā māḥṭtsh Līh Flāsh	If it had a lower price and had a flash, it would have been worth a lot more than this. It is a clear failure from Sony that they didn't add a flash.
3	جهاز جبار ويتفوق علي نظرائة من الايباد والسامسونج بجد رابع بس لم ياخذ حجم الدعايه المطلوبه	Jihāz Jabbār wytfwq ‘Alī nẓrā’h min alāybād wālsāmswnj bi-jadd rāá‘ Bass lam yākhdh ḥajm ald‘āyh almṭlwbh	This an excellent device that outperforms its counterparts from iPad and Samsung which is really great, but it did not receive enough publicity.
4	7 روايات في رواية ، تهكم وسخرية وأدب وأشياء اخرى لن تمل منها	7 Riwāyāt fī riwāyah, thkm wskhryh wa-adab wa-ashyā’ ukhrá lan tml minhā	This book has seven novels in a novel: sarcasm, irony, literature and other things that you will not get bored of.
5	يستاهل لانه دعم اللغة العربية لكن ياريت يثبت البرنامج على رقم المستخدم وليس على حساب	Ystāhl lānh Da‘m al-lughah al-‘Arabīyah lākin yāryt yuthbatu al-Barnāmaj ‘alá raqm al-mustakhdam wa-laysa ‘alá ḥisāb	It is worth it because it supports the Arabic language, but I hope that I can register with my mobile number not my account number.
6	الذي فاجئني وزعجني هو عدم وجود ريموت فيها بالمقارنة مع السعر	Alladhī fāj’ny wz‘jny huwa ‘adam wujūd rymwt fīhā bi-al-muqāranah ma‘a al-si‘r	What shocked and annoyed me was the lack of a remote control in it compared to the price.
7	تطبيق رائع وفيه خصوصيه محدش يعرف دخلت امتي خرجت امتي انت حتي لو بتكتب اللي قدامك مش بيعرف	Taṭbīq rā’i‘ wa-fīhi khṣwṣyh Maḥaddish ya‘rifu dkhlt amty kharajat amty anta ḥattá Law btktb Illī qdāmk mish by‘rf	This is a wonderful application that protects your privacy; no one knows when you logged in or logged out. Even while typing, the person in front of you will not know that you are typing.
8	فكرني بفلاش و سماش والحاجات دي بس مفيش كلمات متقاطعة	Fkrny bflāsh wa smāsh wālḥājāt Dī Bass mfysh Kalimāt mutaqāṭi‘ah	It reminded me of ‘Flash and Smash’ and these things, but this one does not have crossword puzzles.
9	مرررررره حلو وعصيراته فرش ومو حاطين له لاسكر ولا مويه كله فرش	Mrrrrrh Ḥulw w‘ṣyrāth farsh wmw ḥāṭyn la-hu lāskr wa-lā mwyh kullahu farsh	It is extremely delicious, and its juices are fresh, and they do not add sugar or water; it is all fresh.
10	نصيحتي ان في محلات فطاير تانية في حدايق حلوان ممكن تكون أفضل كتير	Naṣīḥatī an fī maḥallāt fṭāyr tānyh fī Ḥadāīq Ḥulwān mumkin takūn afḍal kitīr	My advice is that there are other pastry shops in Hadyaa Helwan that could be much better.
11	اول روايه اقرأها لباولو كويلهو ولااعتقد انها الاخيره	Awwal riwāyah aqr’hā lbāwlw kwylhw wlāā‘tqd annahā alākhyrh	This is the first novel I read for Paulo Coelho, and I don't think it will be the last.
12	لم استسيغ نزار فى شعر الفصحى اعتقدوا فى الشعر الحر اكثر ابداعا	Lam astsygh Nizār fī shi‘r al-fuṣḥá a‘tqdwā fī al-shi‘r al-Ḥurr akthar abdā‘ā	I did not like Nizar in classical poetry. I think in free verse he is more creative.
13	فندق مريح صراحة اسعارة جيدة ما بين 150 الى 300 ريال الليلة ونظيف جدا	Funduq mryḥ ṣrāḥh as‘ārh Jīdah mā bayna 150 ilá 300 Riyāl al-laylah wa-Naẓīf jiddan	It is a comfortable hotel, frankly; it has good prices, between 150 to 300 riyals per night, and it is very clean.
14	أنصح بيه أي حد بيدرس هندسة كهربية أو عايز يدرسها هتوفر عليه جنون كتير	Anṣḥ Bīh Ayy ḥadd bydrs Handasat khrbyh aw ‘āyiz ydrshā htwfr ‘alayhi Junūn kitīr	I recommend it to anyone who studies electrical engineering or wants to study it; it will save a lot for him.
15	ربما يكون الكتاب جيداً ولكن بروز شخصية الكاتب المتملقه تفسد ذلك؟	Rubbamā yakūn al-Kitāb jayyidan wa-lakin Burūz shakhṣīyah al-Kātib almtmlqh tufsidu dhālika?	The book may be good, but the author's fawning character spoils it.
16	ما إن رأيت ولا سمعت بمثله	Mā Inna ra’aytu wa-lā sami‘t bi-mithlih	I have neither seen nor heard of anything like it.
17	ما هذه الرومانسية الحالمة وما هذا الاسلوب الناعم الجميل هذه الرواية من اجمل ما قرات على الاطلاق	Mā Hādhihi al-rūmānsīyah al-ḥālimah wa-mā Hādhā al-uslūb al-Nā‘im al-jamīl Hādhihi al-riwāyah min ajmal mā qrāt ‘alá al-iṭlāq	What is this dreamy romance, and what is this soft and beautiful style? This novel is one of the most beautiful novels I have ever read.
18	أليس هذا بالحق	Alīs Hādhā bi-al-Ḥaqq	Isn't that right?
19	مَا هَٰذَا بَشَرًا إِنْ هَٰذَا إِلَّا مَلَكٌ كَرِيمٌ	Mā haādhā basharan in haādhā illā malakun karīm	This is not a man; this is none but a noble angel **
20	جدة غير	Jiddah ghayr	Jeddah is different/unique.

21	كتاب غير مجرى تفكيرى خلاه اوسع خلانى أثق اووى فى العلامات	Kitāb ghayr majrá tfkyrá khlāh awsaʿ khlāná athq awwwá fī al-ʿalāmāt	This book changed my way of thinking; it broadened my mind and made me trust the signs strongly.
22	تجنننننن انا بصراحه مش بس الفندق جميل كل حاجه زورتها كانت جميله قوي قوي قوي قوي	Tjnnnnnn anā bṣrāḥh mish Bass al-Funduq Jamīl kull ḥājh zwrthā kānat Jamīlah Qawī Qawī Qawī Qawī	Amazing! Not only is the hotel beautiful but also everything I visited there was very very very very beautiful.
23	أجمل ما فيه هو إفطاره	Ajmal mā fīhi huwa ifṭārh	The best thing about it is its breakfast.
24	كتاب خفيف و واقعي و بعيد عن المبالغة تماما كل شيء فيه حقيقي	Kitāb khafīf wa wāqiʿī wa baʿīd ʿan al-Mubālaghah tamāman kull Shayʾ fīhi ḥaqīqī	This is a light and realistic book which is absolutely far from exaggeration; everything in it is real.
25	وَتَرَى النَّاسَ سُكَارَىٰ وَمَا هُم بِسُكَارَىٰ	Watará alnnāsa sukāráā wamā hum bisukāráā	You will see the people [appearing] intoxicated while they are not intoxicated [**]
26	الكلمة قد تفعل فى الانسان ما لم تفعلة الادوية القوية لك كل قدير	al-Kalimah qad tfʿl fī al-insān mā lam tfʿlh al-adwīyah al-qawīyah laka kull qdyr	The effect of a word may be stronger than the effect of medicines.
27	السلعة ليست بالجودة المطلوبة وقد اشتخدمتها لمرة واحدة فقط ولم ارجع لاستخدامها مرة ثانية	Alslʿh laysat bāljwdh al-maṭlūbah wa-qad ashtkhdmthā li-marrah wāḥidah faqaṭ wa-lam arjʿ lāstkhdāmhā marrah thānīyah	The item is not of the expected quality, and I only used it once and did not use it again.
28	جلسات رائعة جلسات المطعم الخارجية رائعة خصوصا في فصل الربيع والشتاء اما الاكل فجيد نوعا ما	Jalasāt rāʾiʿah jalasāt almṭʿm al-khārijīyah rāʾiʿah khṣwṣā fī Faṣl al-Rabīʿ wa-al-shitāʾ amā alākl fjyd nwʿā mā	The atmosphere of the outdoor restaurant is wonderful, especially during spring and winter, but the food is not that good.
29	وَجَعَلَ لَكُمُ السَّمْعَ وَالْأَبْصَارَ وَالْأَفْئِدَةَ ۙ لَعَلَّكُمْ تَشْكُرُونَ	Wajaʿala lakumu alssamʿa wālʾabṣāra wālʾafʾidata ۙ laʿallakum tashkurūn	He [Allah] made for you hearing and vision and intellect that perhaps you would be grateful [**]
30	الَّذِينَ يَظُنُّونَ أَنَّهُم مُّلَاقُو رَبِّهِمْ وَأَنَّهُمْ إِلَيْهِ رَاجِعُونَ	Alladhīna yaẓunnūna annahum mmulāqū rabbihim waʾannahum ilayhi rājiʿūn	Who are certain that they will meet their Lord and that they will return to Him [**]

[*] The transliteration is accomplished by the developed tool at CAMeL Lab, New York Abu Dhabi University (http://romanize-arabic.camel-lab.com/)

[**] The source of translation is King Saud University Mushaf (https://quran.ksu.edu.sa/)