Open Access
Full Length Article

ORIGINAL ARTICLE

A Novel Morphological Rule-based Approach for Urdu Text Sentiment Analysis

Uzma Naqvi[1], Abdul Majid[2], Sana Shokat[3], Adnan Azeem[4]

[1,2,3,4] Department of Computer Science & IT, University of AJ&K

ABSTRACT

Sentiment Analysis (SA) is an exciting research area in Urdu text classification tasks. Due to its complicated morphology, sentiment analysis in Urdu poses various challenges. The lexicon-based technique is the preferred approach for Urdu sentiment analysis for its high accuracy rate. This approach requires a comprehensive manual list of positive and negative words. Sentiment classification can be effectively achieved using morphology-driven methods. For this purpose, an effective Urdu sentiment analyser employing morphological rules for performing Urdu sentiment analysis is proposed. The application of the rules depends on the input words' lexical category and prefixes. Sentiment classification rules are proposed based on the prefixes attached to negative sentiment carrier words. A thorough evaluation with reasonable Urdu text document/dataset size is performed. The empirical assessment proved that the proposed rule-based method achieved satisfactory precision, recall, and accuracy.

Keywords: *Natural Language Processing; Rule-based, Sentiment Analysis; Urdu Language*

Author`s Contribution
[1,2] Data analysis, interpretation, and manuscript writing, Active participation in data collection
[2,3,4] Conception, synthesis, planning of research. Interpretation and discussion

Address of Correspondence
Abdul Majid
Email: majid@ajku.edu.pk

Article info.
Received: July 6, 2023
Accepted: November 13, 2023
Published: December 30, 2023

Cite this article: *Naqvi U, Majid A, Shokat S, Azeem A. A Novel Morphological Rule-based Approach for Urdu Text Sentiment Analysis. J. inf. commun. technol. robot. appl.2023; 14(1):1-07*

Funding Source: *Nil*
Conflict of Interest: *Nil*

INTRODUCTION

Over the last decade, websites have adapted to the changing needs of their users by making their content more accessible and inclusive. By expanding their content to include marginalized languages like Urdu, websites can reach a wider audience and create a more diverse online community. This shift from monolingual to multilingual content has enriched the user experience and helped bridge linguistic and cultural gaps. Urdu is widely spoken in the Indo-Aryan region, and about 170.2 million people use the Urdu language for communication purposes [1]. Urdu is the official language of Pakistan and India's second language used by millions of people worldwide for communicating.

Sentiment Analysis (SA) is recognizing how emotions are written in the text and whether the utterances give a negative (undesirable) or positive (desirable) opinion on the topic. Therefore, sentiment analysis involves identifying emotional expressions, the polarities and

strengths of expressions, and their relationships to topics [2]. There are several approaches to sentiment analysis: lexicon-based, machine learning (ML), and deep learning (DL). Lexicon-based methods make use of a word list and associated word sentiment. This method may be either corpus-based, dictionary-based, or manual. The sentiment score is calculated by taking the summation of the polarity of lexicons in the text if found in the dictionary [3]. A corpus of a list of words with specified polarities and an algorithm are required [4]. ML methods depend upon supervised classification approaches to detect and frame sentiment as positive or negative. However, this approach requires labelled data to train classifiers. Several probabilistic, linear, decision-tree, rule-based, and unsupervised methods can be used for SA.

The legitimacy of the opinions is critical to ensure they are offered from a reliable source. Verification of the opinion holder is another issue, whether they are experts in that field. It is essential to consider the opinion of the appropriate person in the field. There is no standard metric available that can measure the expertise of the opinion holder, but by accessing the profile of the opinion holder, the expertise can be somehow verified. Typically, the opinions are on specific topics and issues. The developed techniques are usually domain-dependent. The sentiments in one domain may appear positive, but the same may seem negative in another domain. Opinion Mining is also complicated because people express themselves in other languages according to their style. In the case of Urdu, some people prefer to use Roman Urdu when providing opinions. Determining an opinion as positive or negative is another issue. There may be words in the opinion that can be used both in a negative and positive sense (i.e. context-dependent words). In linguistics, intensifiers are a lexical category of modifiers that adds an expressive context to the modifier word but do not contribute to the part-speech meaning of the phrase.

Urdu is a widely used language that has yet to receive much attention from the computational linguistics field. Compared to popular English literature centered on adjectival, Urdu literature differs in syntactic, orthographic, and morphological features. Authors [5] claim that new and updated techniques for sentiment analysis should be used. Urdu also has syntax like Hindi but a script dissimilar to Hindi and phonology [6].

To overcome the obstacles in SA, especially in the context of the Urdu language, an approach is proposed based on morphologically motivated rule creation. The proposed approach introduces morphological-based prefix-checking rules for sentiment analysis. This technique involves the identification of negative words through their prefixes. Positive words are those whose prefixes do not match the prefix list, while negative words are those matched with the prefix list. Negative and positive words Intensifiers are used to increase the strength of the sentiment expressed in the words. The study's primary contribution is to lay the groundwork for the design of algorithms that utilize word structures to create sentiment analysis rules. The following are the specific contributions of this work.

- Preprocessing of data Set includes data cleaning and tagging data with Parts of Speech.
- Creation of morphological rules for identifying negative words in the sentence.
- Application of morphological rules for Urdu text on dataset.
- Evaluation using accuracy, precision, recall, and F1 Score.

LITERATURE REVIEW

This section discusses Urdu loan words with negative opinions and datasets and presents the work on text sentiment analysis techniques employed for Urdu in literature.

Root words are those free words that have meaning even when used alone. An inflected/derivative word is formed when affixes are attached to the root word. Such as 'باخبر ' is an inflected form of the word 'خبر'.

Urdu and are both Indo-Aryan languages; however, Urdu is written in Arabic script, and Hindi is written in Devanagari[7]. Urdu vocabulary has its linguistic roots in Sanskrit and for over 75% of and nearly 99% of Urdu verb structure. Hindi adjectives that have the prefix 'ا' may represent negative words such as اپاہج and اکھڑ. In Sanskrit, ان prefix represents negating words that begin with a vowel. In Urdu, it forms words like ان پڑھ and انہونی that carries the negative sentiments. Other negative

prefixes in Urdu that are borrowed from Sanskrit are ن [ni], نر[nir], and نس[nish].

Similarly, there are prefixes, such as 'لا', when attached to a word, which make the antonym form of the word. E.g. 'پرواہ' (Parwah (care)) becomes 'لاپرواہ' (la Parwah, careless).

Morphology and the lexicon connect other linguistic modules like syntax, semantics, and phonology, explained Katamba and Stonham[8]

Mukund and Srihari created an Urdu corpus titled 13 with a semantic role by applying cross-lingual projection [9]. Humayoun et al. have created another corpus-based Urdu vocabulary [10]. Syed et al. create a sentiment-annotated lexicon [11]. Different researchers created rule-based sentiment classifiers using Urdu lexicons representing their carrier sentiments. Muaz et al. created the POS-tagged corpus [12].

Researchers have shown interest in rule-based sentiment classifiers that utilize sentiment carrier words. SentiUnit, the sentiment carrier expression, is extracted and categorized based on intensity and orientation. Syed et al. created an Urdu sentiment-annotated lexicon with this lexicon score [13]; in a separate study, they tested their model with two corpora of reviews, one in the area of movies and one in the domain of electronics, and reached 74% accuracy. In another effort by Syed et al., targets are coupled with SentiUnits to improve performance by up to 82.5 percent accuracy. In this context, the noun phrases about which the opinions are formed are called targets[14]. Daud et al. extracted adjectives and compared them to an opinion word lexicon created by hand to determine the polarity of opinions. In their method, 21.1 percent of opinions are classified incorrectly [15].

Reman et al. [16] introduced a framework for sentiment analysis in Urdu comments. The lexicon-based architecture assigns polarity to the tokens produced by Urdu sentences. The lexicon has 7335 items, 2607 of which are harmful and 4728 of which are positive. The overall polarity of a sentence is the total weight of all the words in the sentence. Experimentation using a data set comprising 124 Urdu comments from various Urdu websites is carried out to assess the performance of the suggested framework. The architecture has a 66 percent total efficiency.

For Urdu sentiment analysis, a lexicon-based technique is utilized by Mukhtar et al. [17]. Aside from the standard technique of having negations, adjectives, and nouns, authors also incorporated context-dependent items, intensifiers, and verbs in the lexicon. They also experimented with supervised methods to classify Urdu text[18] Asghar et al. [19] calculated sentiment scores of Urdu language lexicons by translating them into English and acquired their sentiment score. The authors also considered modifiers for sentiment classification.

A hybrid sentiment analysis model is proposed[20] that uses dependency-based grammatical rules with deep learning(DL) models to identify sentence polarity. The performance of the deep learning model is improved when applied along with grammatical rules.

In a study [21], the authors use traditional machine-learning approaches to classify proverbs and idioms and leverage linguistic aspects of the Urdu language for sentence-level sentiment analysis. They stated that the J48 classifier exhibits superior performance in sentiment classification, with 90% accuracy and an F-measure of 88%.

Lexicon-based sentiment classifiers outperformed the machine learning methods[21]. As the Urdu language has an extensive vocabulary, building sentiment lexicons for it is a laborious task. Therefore, there is a pressing need to find a more efficient way that requires fewer resources to perform sentiment analysis on Urdu text. This study proposes a rule-based sentiment classification method that uses prefixes to classify sentiments. The goal is to assist in making a sentiment analysis process for the Urdu language. Efficient sentiment analysis for Urdu text is challenging due to its vast vocabulary. To simplify this task, a rule-based method that utilizes prefixes for sentiment classification has been proposed in this study.

RESEARCH METHODOLOGY

Urdu contains loan words from various languages, including Arabic, Persian, Hindi, and Turkish. In the Urdu language, the words that start with “Alif (الف)”, “An (ان)”, “bay (ب)", “Noon (ن)”, “Nar (نر)|”, “Kaaf (ک)”, and “Bin (بن) borrowed from the Hindi language have adverse meaning. In this study, these prefixes are exploited to prepare rules for identifying negative words in the

sentence. A prefix list of these characters is prepared to identify negative words. Rules related to the prefixes are given in Table 1.

Table 1. Examples of Prefixes and their Related Rules

Letter (حرف)	Words (الفاظ)	Sentences (جملے)
Rule 1: If the word starts with ("بے") THEN the word is Negative.		
بے	بےدھڑک	ہم گناہوں کی طرف بےڈھرک دوڑے چلتے ہیں۔
	بے حس	دنیا اسرائیل کے مظالم پر بےحس ہے۔
	بےجوڑ	بےجوڑ رشتوں کی اہمیت کم ہوتی۔
Rule 2: If a word starts with ("ن") then the word is a Negative word.		
ن	نکمّا	نکما انسان کاہلی کی ایک بڑی مثال ہے۔
	نفرت	سوشل میڈیا نفرت آمیز مواد سے بھر گیا ہے۔
	نامراد	کون نامراد رہے
Rule 3: If a word starts with (" ا") then the word is a Negative word.		
ا	ابتر	فلسطین کے حالات ابترہیں
	افسوسناک	پاکستان کے معاشی حالات افسوسناک ہیں۔
Rule 4: If a word starts with ("ک") then the word is a Negative word.		
ک	کمتر	شہرمیں رہنےوالےگاؤں مینررہنےوالوں کو عُمُوماً کمترسمجھتے ہیں۔
	کراہ	مریض درد سے کراہ رہے ہیں۔
Rule 5: If a word starts with ("بن") then the word is a Negative word.		
ن	بن بلائے	بن بلائے مہمان بلاءَ جان ہوتے ہیں۔
	بن باس	کسی کے ساتھ رہنا اور وہ بھی اجنبی بن کرکسی کو کیا پتا کتنا بڑا بن باس ہوتا ہے

Figure 1 describes the steps proposed for sentiment analysis using the flow model. In the first step, the dataset is preprocessed. Stopwords, numbers, and punctuation marks are removed. The cleaned data is tagged with Parts of Speech information.

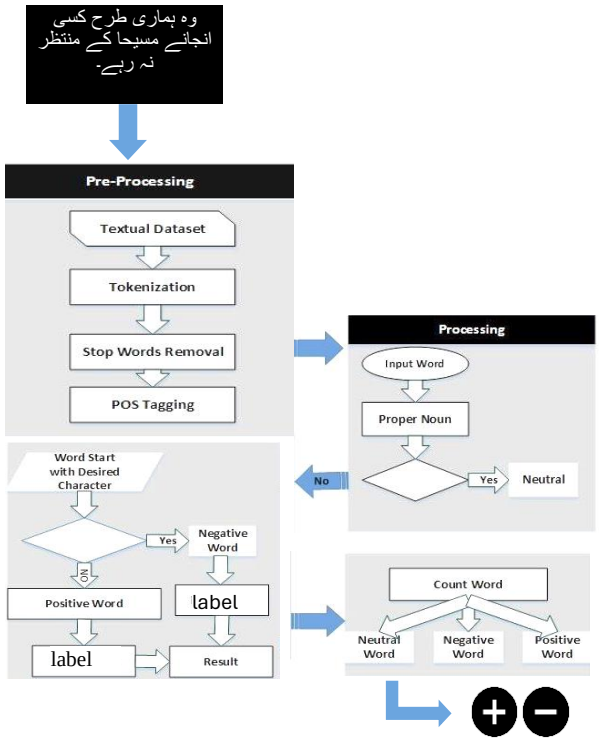

Figure 1. Morphological rule-based Urdu Architecture

Preprocessed and tagged data is assessed for its lexical category. If the word is a noun/pronoun, it is declared neutral and filtered from being further processed. In other cases, Prefixes are extracted from remaining words and compared with characters in the prefix list. The words contain the same prefixes as in the prefix list for negative words and are identified as negative; otherwise, the word is declared positive. Algorithm 1 describes a comprehensive set of procedures for performing sentiment analysis on a text. The final step involves computing the total count of positive, negative, and neutral words within a given sentence. The sentence's sentiment is determined by considering the predominant sentiment in the sentence. A sentence is considered to have a positive sentiment if its word count is mostly composed of positive words. On the other hand, it would be categorized as having a negative sentiment if the majority of the words in it are negative. The sentence can be categorized as neutral when the sentiments are evenly distributed.

Algorithm-I:

Input: POS tagged Sentence **S**

Output: **S** with Polarity associated

1. For each word ***w*** in **S**
   If ***w*** is a **Noun/Pronoun,** then.
   Mark it as Neutral
   NU++
   Else ***p***=Get_Prefix(***w***)
   If Found (p, prefix_list) then
   Mark it as Negative
   NG++
   Else
   Mark it as a Positive
   PV++
2. If NU> NG and NU> PV then
   S is Neutral
   Else IF NG>NU and NU> PV then
   S is Negative
   Else
   S is Positive
3. End For

RESULTS AND DISCUSSION

This section describes the dataset and experimental results of the proposed methodology on the selected dataset. Python version 3.12 is used to implement the methodology in the Spyder notebook. Stanza [22] is used to tag the data with parts of speech.

Data Set: This study used a dataset comprised of 3047 sentences in CSV format. Data is collected from different genres like current affairs news, blogs, sports news, analysis, and social websites.

Evaluation Measures: To prove the effectiveness of the proposed methodology, accuracy, F-1 measure, precision, recall, and AUC are used as the evaluation metrics.

Results: Preprocessing is performed before the application of rules. It reduces the data size to be analyzed substantially after removing the stop words. Reduction in the size of data accelerates the analysis process.

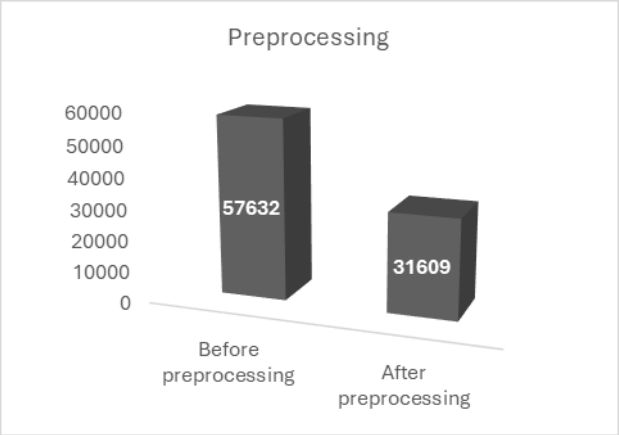

Figure 2. Data Statistics After Preprocessing

The experiment's outcome was recorded as a confusion matrix (Figure 2) and an accuracy graph (Figure 3). It can be observed that the model classified the negative and positive sentences of the dataset. Models successfully classified correctly more positive instances than negative instances. The low true negative rate is due to the identification of negative words solely based on prefixes. Figure 5 demonstrates the effectiveness of the proposed methodology by achieving significant accuracy and an F-1 score.

TARGET / OUTPUT	Neg	Pos
Neg	60 30.000%	29 14.500%
Pos	40 20.000%	71 35.500%

Figure 3. Confusion Matrix

The proposed model attained an excellent AUC score that represents that the model can classify negative and positive sentences effectively.

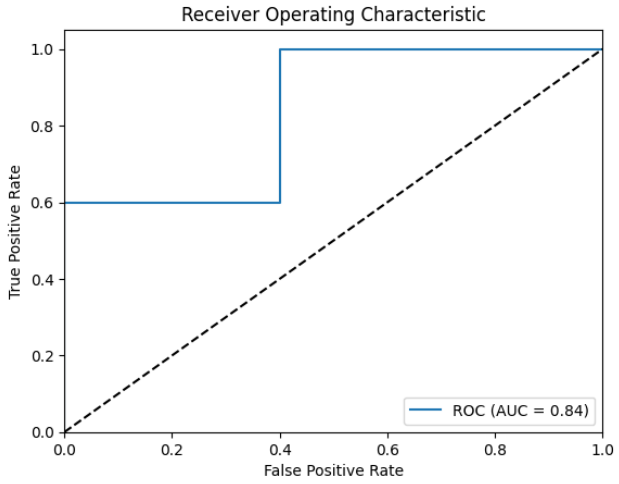

Figure 4: ROC Achieved by Proposed Model

This study aimed to develop a framework to identify negative words and categorize sentences into positive or negative classes. To achieve this, the study utilized prefixes on loan words that express negative opinions. The model correctly predicted the polarity of the sentences that fulfilled the criteria, as shown in Table 2.

Table 2: Qualitative Analysis of Model's Performance

Letter	Word	Sentences	Actual	Predicted	Letter
ا	امر	ہم زندگی اس طرز میں بسر کر رہے ہیں جیسے ہم امر ہیں	Neg	Neg	ا
بے	بےجوڑ	معاشرے کی بےحسی کا یہ عالم ہے کہ بےجوڑ رشتوں کی اہمیت کم ہوتی۔	Neg	Neg	بے
ن	نکما	نکما انسان قابلی کی ایک بڑی مثال ہے۔	Neg	Neg	ن

However, as we have only used a limited dataset, further experimentation with more data is necessary to establish a strong foundation. As w As Urdu contains loan words from Persian and Arabic, it is crucial to explore comparable patterns in these language words. We did not cover negation in this study, but it is essential as it can change the meaning of a sentence.

CONCLUSION

This study presents an investigation of a morphological rule-based technique for sentiment analysis (SA) in Urdu text. The proposed method uses morphological rules based on lexical categories and prefixes for binary classification (Positive, Negative). Words other than nouns/pronouns are considered for the sentiment evaluation through prefixes. The rule-based sentiment classification approach has proven to be more effective for languages with complex morphology like Urdu, and it provides a better approach for extracting sentiments in the Urdu language according to grammar and linguistics. Through experiments, it has been shown that this approach has better performance for sentiment analysis for Urdu text.

Additional research is required to identify comparable patterns that can decrease the reliance on manual lexicon formation. To calculate the sentiment of a sentence, the technique can be enhanced by factoring in sentiment scores that include intensifiers and negation.

REFERENCE

[1] U. Naqvi, A. Majid, and S. Ali Abbas, "UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods," IEEE Access, 2021, doi: 10.1109/ACCESS.2021.3104308.
[2] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the 2nd International Conference on Knowledge Capture, 2003, pp. 70–77.
[3] N. Mukhtar and M. A. Khan, "Effective lexicon-based approach for Urdu sentiment analysis," Artif Intell Rev, vol. 53, no. 4, pp. 2521–2548, 2020.
[4] N. Mukhtar and M. A. Khan, "Effective lexicon-based approach for Urdu sentiment analysis," Artif Intell Rev, vol. 53, no. 4, pp. 2521–2548, 2020.
[5] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Sentiment analysis of Urdu language: handling phrase-level negation," in Mexican International Conference on Artificial Intelligence, 2011, pp. 382–393.
[6] Y. Dubinsky, T. Catarci, S. R. Humayoun, and S. Kimani, "Integrating user evaluation into software development environments," in 2nd DELOS conference on digital libraries, Pisa, Italy, 2007.
[7] Vasudha Dalmia, Hindu Pasts. Sunny Publisher, 2017.
[8] F. Katamba and J. Stonham, Morphology: Palgrave Modern Linguistics, 2nd ed. Macmillan Education UK, 2006.

[9] S. Mukund and R. K. Srihari, “Analyzing Urdu social media for sentiments using transfer learning with controlled translations,” in Proceedings of the Second Workshop on Language in Social Media, 2012, pp. 1–8.
[10] M. Humayoun, R. M. A. Nawab, M. Uzair, S. Aslam, and O. Farzand, “Urdu summary corpus,” in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), 2016, pp. 796–800.
[11] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, “Lexicon-based sentiment analysis of Urdu text using SentiUnits,” in Mexican International Conference on Artificial Intelligence, 2010, pp. 32–43.
[12] A. Muaz, A. Ali, and S. Hussain, “Analysis and development of Urdu POS tagged corpus,” in Proceedings of the 7th Workshop on Asian Language Resources (ALR7), 2009, pp. 24–31.
[13] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, “Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text,” Artif Intell Rev, vol. 41, no. 4, pp. 535–561, 2014.
[14] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, “Associating targets with SentiUnits: A step forward in sentiment analysis of Urdu text,” Artif Intell Rev, vol. 41, no. 4, pp. 535–561, 2014, doi: 10.1007/s10462-012-9322-6.
[15] M. Daud, R. Khan, A. Daud, and others, “Roman Urdu opinion mining system (RUOMiS),” arXiv preprint arXiv:1501.01386, 2015.
[16] Z. U. Rehman and I. S. Bajwa, “Lexicon-based sentiment analysis for Urdu language,” in 2016 Sixth International Conference on innovative computing technology (INTECH), 2016, pp. 497–501.
[17] N. Mukhtar and M. A. Khan, “Effective lexicon-based approach for Urdu sentiment analysis,” Artif Intell Rev, no. 0123456789, 2019, doi: 10.1007/s10462-019-09740-5.
[18] N. Mukhtar and M. A. Khan, “Urdu Sentiment Analysis Using Supervised Machine Learning Approach,” Intern J Pattern Recognit Artif Intell, vol. 32, no. 02, p. 1851001, 2018, doi: 10.1142/S0218001418510011.
[19] M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, and S. Ahmad, “Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language,” Expert Syst, vol. 36, no. 3, pp. 1–19, 2019, doi: 10.1111/exsy.12397.
[20] U. Sehar et al., “A hybrid dependency-based approach for Urdu sentiment analysis,” Sci Rep, vol. 13, no. 1, p. 22075, Dec. 2023, doi: 10.1038/s41598-023-48817-8.
[21] N. Mukhtar, M. A. Khan, and N. Chiragh, “Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains,” Telematics and Informatics, vol. 35, no. 8, pp. 2173–2183, 2018, doi: 10.1016/j.tele.2018.08.003.
[22] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A {Python} Natural Language Processing Toolkit for Many Human Languages,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020. [Online]. Available: https://nlp.stanford.edu/pubs/qi2020stanza.pdf