ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 2, pp: (428-436), Month: April - June 2019, Available at: www.researchpublish.com

# Developing a Normalizer for San'ani Arabic Social Media Texts

Mohammed Sharaf Addin

CAS in Linguistics, Osmania University, Hyderabad, India.

Department of English, Faculty of Arts, Thamar University, Yemen.

Author email id: ma.alshami22@gmail.com

***Abstract:*** **The successive and mounting flow of data that is generated by social media platforms everyday raises a question of the validity of such data and to what extent can such data be used by machines. One of the most challenging problems with social media data is the inconsistency of texts which is influenced by many factors. Such a product appears mingled with many of ill-formed data which are problematic for machines and natural language processing tasks. The primary aim of this paper is to identify and discuss different errors generated by the social media text of San'ani Arabic and hence, develop a text normalizer for correcting and standardizing such errors. The identification process is performed manually on a corpus of 158,279 tokens and 20,000 types from texts that are extracted from Facebook and Telegram platforms. As a result, 64,040 tokens and 1,741 types with errors are identified. These errors were classified into two broad categories (i.e., tokenization as well as typographical) and about 15 sub-types based on their frequency as well as typology. Further classification is made based on regularities of these errors. A rule-based as well as dictionary-lookup normalizer is developed using python programming and Django API which shows 99% performance among tokens and 98% among types.**

***Keywords:*** **Social Media Texts, San'ani Arabic, Normalizer, Error types, Error Identification.**

## I. INTRODUCTION

One of the most recent data sources in the field of Natural Language Processing (NLP) is social media platforms. Working with Social media data has its own challenges and requirements due to the peculiar nature of such data. Social networks data is highly non-standardized [1], [2], [3]. It is characterized with the use of emoticons, shortening, non-standard spellings, lengthening and omission of some letters [2], [4]. These characteristics make social media data unsuitable for NLP applications. To overcome this obstacle, a pre-processing step is required in which data is normalized.

Normalization is the process of identifying linguistic noise and reducing it. For social media text, shallow normalization is conducted due to non-standardization [2]. Working on Arabic text necessitates pre-processing due to the language nature. According to Habash [5] Arabic orthographic normalization has its own issues that have to be addressed. He discussed different types of Arabic orthographic normalization such as lexical/letter normalization. Buckwalter [6] discussed the challenges of Arabic orthographic variation, distinguishing two types namely normal variation which refers to human perception of what is permissible and mechanical variation which refers to computer input symbols.

Dialectal Arabic normalization has its own obstacles and challenges. The reason behind this is related to the diglossic nature of Arabic where two varieties are used by the speakers. There is a high variety which is Modern Standard Arabic (MSA, henceforth) that represents the formal language of media, education and is the written form. The low variety is the dialect which is the informal language of everyday communication and is mostly spoken. Fortunately, Arabic speakers started writing in their informal dialects when communicating through social media platforms. Obtaining dialectal data from social media applications is recently considered a very rich source of data though it poses its own challenges in

comparison with MSA data. In this work, we target San'ani Arabic (SA, henceforth) data which is collected and extracted from social media platforms. The normalization task for such data is essential before any further NLP processing.

This paper tries to introduce an original work by developing our normalizer that deals with corrupted social media texts as well as introducing our methodology of error identification. We focus on the task of normalization of San'ani social media data which is extracted from Facebook and Telegram platforms. The aim is to provide an in-depth analysis of the error types and convert out of vocabulary (OOV) words to their in vocabulary (IV) equivalents making non-canonical data standardized and suitable for machine use.

This paper is structured into the following 9 section headings: 1)Introduction; 2) Related work; 3) San'ani Arabic Text Selection; 4) Error Identification Process; 5) Types of Errors; 6) Description of our Normalizer; 7) Results and Evaluation; 8) Conclusion; and References.

## II. RELATED WORK

Though our work deals with San'ani social media text normalization, few works have addressed dialectal Arabic text normalization. In this section, we will concentrate on some researches that are conducted on social media text pre-processing.

For the task of text normalization, researchers adapted different methods to achieve their goals. Some of the proposed methods are rule-based, cognitive-based, word/character level machine translation and neural methods. Such methods work on word level, character level, or both. Liu et al. [1] developed a cognitive driven normalization system that models three human cognitive abilities which are the enhanced letter transformation, visual priming, and string/phonetic similarity. The system deals with the word level and message level claiming accuracy increase of 10% compared to the state-of-the art for English language.

Lusetti et al. [3] adapted neural encoder-decoder (ED) models to normalize Swiss German What's up messages. They claim improvement over the performance of the state of the art character-level statistical machine translation (CSMT). It works on both word and character level. On the other hand, CSMT was earlier used as a method of translation between related Languages such as Catalan to Spanish [7]. Due to its impressive performance on out of vocabulary (OOV), it was used as a method of normalization [8]. Samardzic et al. [8] used an optimal combination of three methods word-by-word mappings, character-based machine translation, and language modeling to normalize Swiss dialects of German.

[9] worked on Turkish language non-canonical variety to normalize text for further NLP processing. They acknowledged the fact that morphologically rich languages need to integrate different methods of analysis and solution to overcome language complexity. The system is built basically into two stages. The first stage analyzes the input to separate the ill-forms and the second stage provides an in-depth analysis and generation through seven layers of error analysis and correction.

For Arabic language, as a morphologically rich language, the adaption of any normalization system has to consider language complexity and lack of annotated data. Neural models have been adapted recently to work on text normalization. For example, Watson et al. [10] applied sequence-to-sequence models to Arabic, but they had to combine other methods to allow some morphological information into the task. This relates to the nature of the language and the lack of annotated data. The rule-based method is usually used for Arabic in combination with other methods to achieve high accuracy. Nawar [11] for instance, built a hybrid error correction system to improve the statistical results using syntactic rules to increase the accuracy level. The work of Arabic normalization is mostly directed to classical or traditional text which is quite different from the dialectal social media text.

As far as our knowledge concerns, there is no any previous normalization system for San'ani Arabic recently available social media texts. Even available normalizers for Arabic are developed for the standard Arabic variety (i.e., MSA). However, there is less focus on the dialectal variety as a result of paucity of sources as well as most people still look at it as a low variety and hence, absence of NLP applications.

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 2, pp: (428-436), Month: April - June 2019, Available at: www.researchpublish.com

## III. SAN'ANI ARABIC TEXT SELECTION

The data used to be examined is selected from SA corpus that is collected from Facebook and Telegram applications. Our selection is based on the most frequent words in the corpus. Lancsbox tool was used for statistical calculations. Word counts are sorted in a descending order (i.e., from the most occurring to least occurring) which range between (2,987 as highest –10 as lowest). Words with less than 10 frequent times were not included. We examined about 158,279 tokens and 20,000 types. Then, using our knowledge based experiment; we tracked all the types manually and extracted the non-canonical, typographical erred words. All extracted erred words are stored in a database and assigned their corresponding standard forms. Ratio between correct and erred words among the corpus is listed as (94,239: 64,040 tokens which is 10 : 7 ) and (18259:1741 types that is equal to 10 : 1) as shown in Fig.1.1.

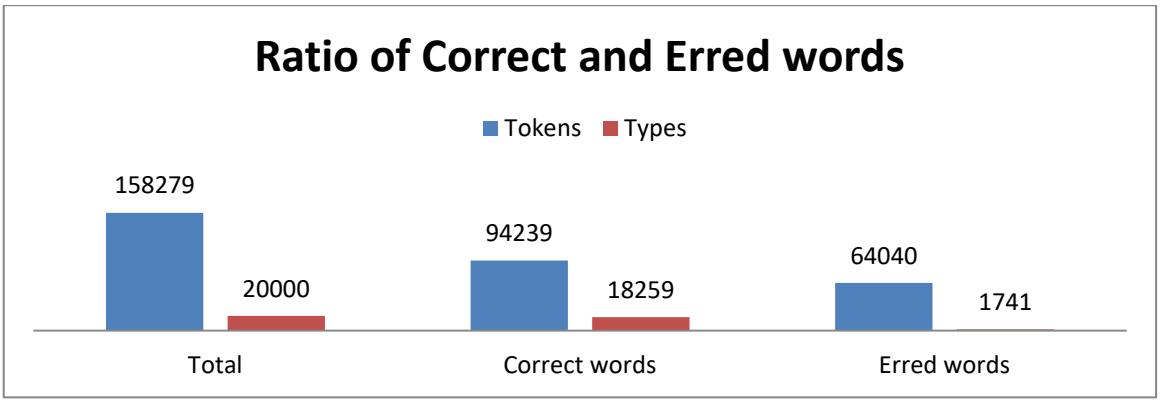


**Fig.1.1 Shows the ratio of valid to invalid words**

## IV. ERROR IDENTIFICATION PROCESS

The identification of errors is performed manually by tracking the sample text. The process of error identification as shown in Fig.1.2 starts by selecting a sample raw text for San'ani data that is extracted from Facebook and Telegram apps. This text is then analyzed statistically using Lancsbox system for tokenization and frequency count. We ended up with a total of 158,279 tokens and 20,000 types. Fig.1.1 shows the statistical details. Then, based on our knowledge experiment as speakers of the dialect, we tracked our corpus manually and extracted all ill-constructed words. The process separates Error-Free words from corrupted or ill-formed words and each is stored separately in a database. The normal and valid words were used as Gold Standard for testing our algorithm. As our purpose is to tackle the ill-formed data, we identified 15 different types of errors. These errors are classified as per types into a) tokenization-based errors (i.e., *AlphaN, LexC,* and *Clitics*; see Table 1.1 for decoding); and b) typographical errors with about 2 subcategorized branches (i.e., social media specific and Dialect specific errors). Social media Specific errors include *LL, Abbrev, SpelE, and DFW*. Dialect specific errors include 8 types (i.e., *HPAP, HaTa, YRE, LAiLA, Dhadh, AID, PhonE* and *aAY*) (Again information decoding can be looked up on Table 1.1).

**TABLE 1.1 Illustrates all types of errors with their codes**

| Label | Code | Explanation |
|---|---|---|
| Letter Lengethenign (elongation) | LL | This includes all words with letter duplication. e.g., قوووى *qawwwwa* 'please'; أمممم *'ummmmm* 'Humming'; هههه *haaaa* 'lol'. |
| Hamza Presence or Absence and Positioning | HPAP | This includes all words with *hamza* problems. e.g., ايمان *'īmān* 'Faith' or *'ayyimān* 'oaths'. |

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 2, pp: (428-436), Month: April - June 2019, Available at: www.researchpublish.com

| | | |
|---|---|---|
| Haa and Taa Marbu:ta | HaTa | This type composes all words ending with *Haa* or *Taa Marbu:ta*.<br>e.g., الدبله *addublah* instead of الدبلة *addublat* *'wedding ring'*; مدرسه *madrasah* instead of مدرسة *madrasat* 'school'. |
| Lexical Concatenation | LexC | This includes all words with lexical tokenization problems.<br>e.g., وقدهي *waqadhi* instead of وقد هي *waqad hi* *'and may she'*; لوتبسري *lawtabsiri* instead of لو تبسري *law tabsiri* 'if you see'; هياناهي *hayyanaahiy* instead of هيا ناهي *hayya naahiy* 'come on'. |
| Derived Foreign Words | DFW | e.g., البارت *al-part* 'the part; ثانكس *thanks* 'thanks'; كنسلت *cansalat* 'she canceled'; بروفايلي *brufaayli* 'my profile'; للتيتشر *li-tishirt* 'for the T-shirt'. |
| Laam ʾalif for illa | LAiLA | This includes لا *laam ʾalif* letter 'not' that is mistaken for the preposition إلى *'ila* 'to'.<br>e.g., لاعندهم *laa'indahum* literally 'not where they are' instead of إلى عندهم *'ila 'indahum* 'to where they are'. |
| Arabic-Indic Digit | AID | This includes all Eastern Arabic-Indic Numbers (i.e., ٠, ١, ٢, ٣…etc.) that are used instead of Western Arabic digits (i.e., 1,2,3,4…..etc.)<br>e.g., ٢٠١٦/١١/٧ "2016/11/7", ١٢:٣٠ " 12:30", ٥ "5",٪٩٢.٥ " %95.5" |
| Interchange of ʾalif maqsu:ra, bare ʾalif and Yaa | aAY | This includes all the words having *ʾalif maqSu:ra, bare ʾalif* and *Yaa* used interchangeably.<br>e.g., خبلا *or* خبلى *khablaA 'silly*, علا or *علىʕalaA 'on'.* |
| phonological Errors | PhonE | This includes all words exhibiting phonological change or errors.<br>e.g., مشو *mashaw* 'not he/it'. |
| Dhaa: ʾalif and Dha:d | Dadh | It includes all words with *Dha:d* or *dhaa ʾalif* error.<br>e.g., الظغط *aldhaghaT* instead of الضغط *alDaghaT* ' the pressure or stress', مواظيع *mawa:dhi:'e* instead of مواضيع *mawa:Di:'e* 'matters / topics'. |
| alphanumeric | AlphN | The problem of alphanumerical error is a tokenization problem.<br>e.g., البنت٢ *albint*٢ 'girl2'; ٩:٠٠الساعه *als:ʕah9:00* 'the time 9:00' . |
| Yaa Letter Reduction | YLR | This includes all words where medial long vowel *Ya:* ي *i:* letter is replaced by short vowel (diacritic) ِ i.<br>e.g., غر *ghir* instead of *ghai*:r 'but'; زد *zid* instead of *zi:d* 'in addition to'; حن *Hin* instead of *Hi*:*n* 'when'; اش *?aish* instead of *?ai:sh* 'what'. |
| clitics | Clitics | This includes all the clitics in the corpus that lose their meanings as a result of wrong tokenization.<br>e.g., بش *bish* it occurs as مابش *ma:bish* 'there is nothing', بحد *biHad* it occurs as *ma:biHad* 'none is there'. |
| Spelling Errors | SpelE | This includes all the words with wrong spellings.<br>e.g., كولنا *kau:lana:* instead of كلنا *kulana:* 'all of us', دقايق *daqa:yiq* instead of دقائق *daqa:?iq* 'minutes'. |
| Short forms or abbreviation | Abbrev | This includes all abbreviations and short forms.<br>e.g., ع ʕ a short form of ʕala 'on or about', ص *S* a short form of صباحاً *Saba:Han* 'A.M.' or 'morning'. |

Fig.1.2 below shows the flow diagram of our methodology for error identification. The proposed methodology helps in diagnosing the most occurring types of errors in the SA social media texts.

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 2, pp: (428-436), Month: April - June 2019, Available at: www.researchpublish.com

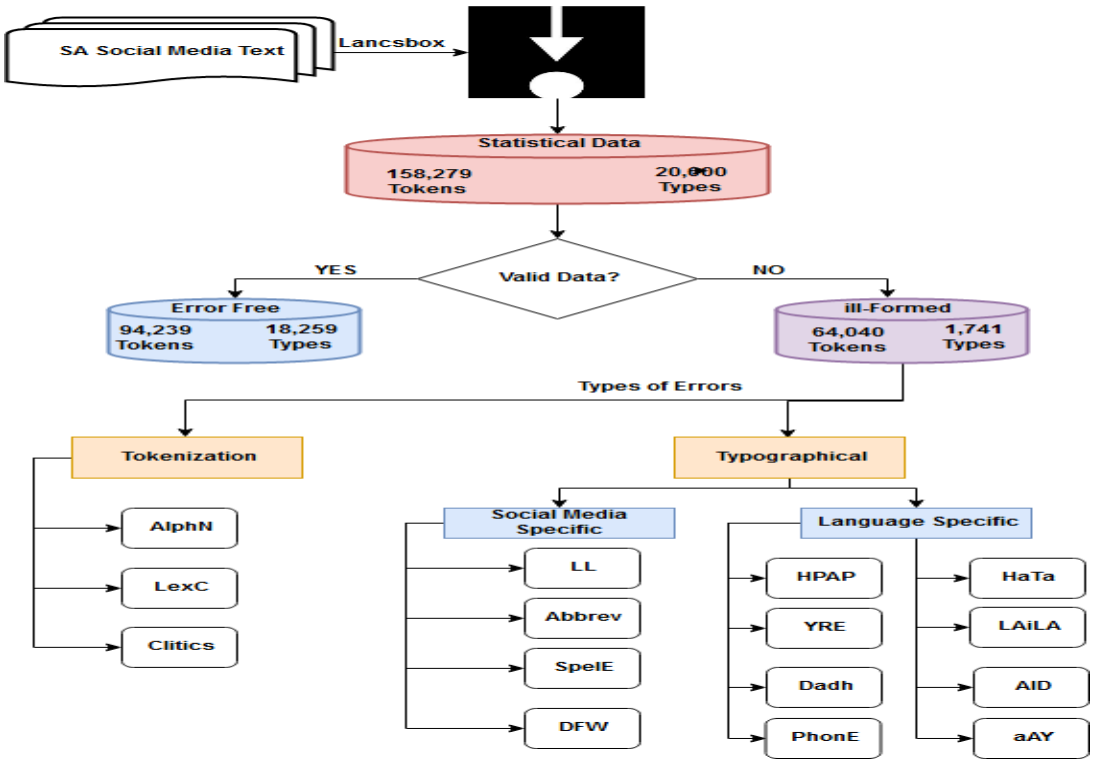


**Fig 1.2 Elucidates Error Identification Flow Diagram**

## V. TYPES OF ERRORS

Table1.1 shows different types of errors that were identified as the most occurring errors in the corpus. The table is divided into three columns. The first column lists all labels of errors; the second assigns corresponding codes for each label for the purpose of paper presentation and ease of information referencing. The third column provides more illustration with given examples.

Our classification approach is influenced by Buckwalter [6] and shares all types of variations of MSA. [6] classified variation into two types: normal and mechanical variations. Normal variation refers to the one which is considered orthographically correct or acceptable; whereas, mechanical variation refers to the ambiguity of typists or writers in choosing the right keyboard input characters. For SA, a more detailed and specific classification of common errors is provided as per the dialectal and environment nature. However, in both MSA and SA, the crucial problems are orthographical in nature. These types of errors are illustrated in detail with examples and their proposed solutions as follows:

Letter Lengthening or elongation is a common social media error. Sometimes repetition of letters is intended for emphasizing some feelings that may be best expressed with the word 'very'. For example, قوي *qawiy* 'strong' and قووي *qawwwiy* may be understood as 'very strong'. Most of Letter repetition is noticed among the long vowel letters. This kind of error is normalized by reducing all letter duplications into a minimum of two letters for all occurrences.

Another type of error which is shared with MSA is described as presence or absence of *hamza* ء (U+0621) and its positioning on some vowel letters. This type of error is noticed as the most frequent one in the analyzed corpus as is shown in Fig. 1.3. The letters include above or below ا *ʾalif* (U+0627), above و *waw* (U+0624) or on ى *ʾalif maqṣūrah* (U+0649). These types of *hamza* positioning are called as normal orthographic variation [6] because they are perceived as orthographically correct or at least acceptable and have only one interpretation.

The other type of orthographical variation is called mechanical variation [6]. An example of this variation is the word اكرم *akram* where the writer is confused in positioning *Hamza*. If *Hamza* is placed above the ا *ʾalif as* أكرم *'akram* the meaning refers *to 'name of a person',* whereas, if *Hamza* is positioned under the *ʾalif as* إكرم *'iikram*, it indicates an imperative

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 2, pp: (428-436), Month: April - June 2019, Available at: www.researchpublish.com

verb meaning 'honour' or 'treat politely'. Same problems are with absence or presence of *maddah* above *ʾalif* آ *ʾalif maddah (*U+0653).

The *Ta: Marbu:Ta* (ة U+ 0629) is another observed normal typographical error in the social media text where writers are observed in most cases to replace it with *ha:* (U+0647) (i.e., *ta: marbu:Ta* without the two dots). Another typographical error is the misspelling of *ʾalif maqṣūrah* (ى U+0649) with ( ي U+064A) as well as final bare *alif (* ا U+0627). Such proper positioning of the mentioned above normal orthographic variation cases is crucial especially for word look-up in the dictionary or lexicon, disambiguation as well as proper transcription.

Lexical concatenation (LexC), Alphanumeric words and clitics are other types of errors that result of improper tokenization. However, this type of error requires proper tokenization for lexical identification and hence, proper categorization. The clause, for example, لوتبسري *lawtabsiri* 'if you see' must be tokenized as *law tabsiri*. This type of tokenization errors affects the statistical calculations and hence, the proper analysis. Another example of this type is tokenization of letters of the same word where a word is mistakenly broken by unintended space. These types of errors are solved by adapting many-to-one approach.

Derived Foreign Words (DFW) is another problematic variation where social media users tend to apply Arabic inflections for some foreign words. This type of error is solved by assigning a single standard spelling of foreign words in the derived cases. Then we add them to the Arabic lexicon and apply Arabic derivational rules on them. The DFW كنسلت *cansalat* 'she canceled' is an example where Arabic feminine marker *Teh* t ( ت U+062A) is attached to the verb كنسل cansal 'cancel' to indicate the past.

In case of numerical representation in the corpus, we observed inconsistency of different format representations of date, time or numerals. The majority of users adapt Western Arabic Digits (WAD) (0-9 U+0030-U+0039). However, some still use Eastern Arabic-Indic Digit (AID) (٠-٩U+0660-U+0669). In such variation, all EAIDs are standardized into WADs.

Another error is restricted to the letter *La:m* + *'alif* 'not' (لا U+0644 + U+0627) which is a negation particle in MSA. However, in the social media text, it is used as a preposition إلى *?ila* 'to' as well as a negation particle لا *la:* 'not'.

Interchange misspelling of the most confusing letters *Da:d* ( ض U+0636) with dhah ( ظ U+0638) is another error which is referred to as (Dadh error). In this case, specific words are assigned each letter and stored in the database. Based on that, suggested normalization is performed.

Dialectal phonological effect results in producing some words which are written as those in the standard variety but have a different meaning. These are expressed as types of words with phonological errors (PhonE). The dialectal word, for example, مشو *mashaw* 'It is not (he)' may be ambiguous with the same MSA word that is also a part of the dialect and means 'they walked'. Phonological type of error is also influenced by the person idiolect, and hence, variations of same form occur. In normalizing these types of errors, we standardize all variants into a single canonical form. YLR error belongs to this phenomenon.

Spelling error (SpelE) is another common error in the social media text which may occur mistakenly or influenced of user's educational background.

Abbreviations or short forms are presented as a single letter which can be ambiguous. These kinds of short forms are standardized into their full forms as per their occurrences in the corpus.

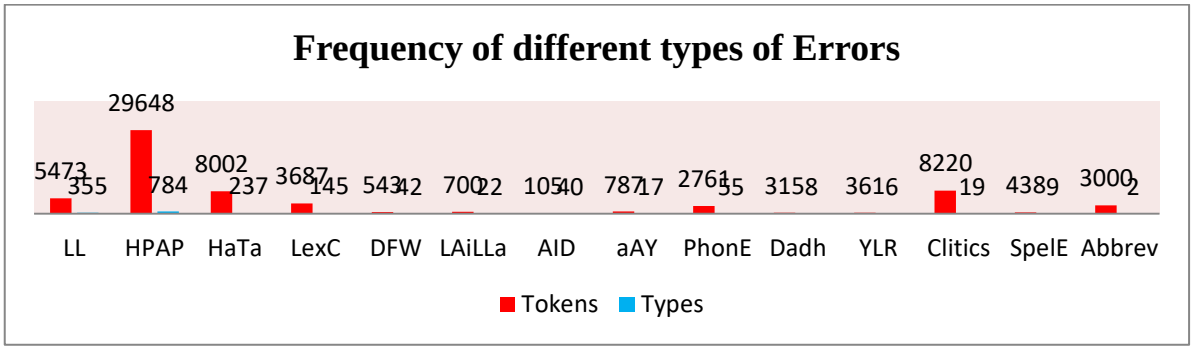


**Fig 1.3 Shows the Frequency of the most types of Errors**

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 2, pp: (428-436), Month: April - June 2019, Available at: www.researchpublish.com

As illustrated in Fig. 1.3, the frequency of different types of errors among the word tokens and types. HPAP is the most frequent type of errors with about 784 types and 29,648 tokens, whereas DFW is noticed as the least common error with 543 tokens and 42 types. Some types of errors are seen to be frequent among tokens whereas other types are observed less. *Abbrev* type of error is an example of this with about 3,000 tokens; however, types are only among two words. The total average among tokens and types is 8,539 and 232 consequently.

## VI. DESCRIPTION OF OUR NORMALIZER

The major aim for designing our normalizer is to tackle all the identified issues in the selected corpus. These issues have been noticed while developing the San'ani social media corpus. Our algorithm is developed using python programming and Django API Framework. The simple description of our normalizer is a hybrid based normalizer that uses both rule-based technique as well as dictionary lookup. Our methodology starts by scrutinizing a random corpus of 158,279 tokens and 20,000 types from San'ani social media texts. This corpus was handled manually by us as linguists and speakers of the dialect. The purpose is to filter out the corpus into valid versus invalid tokens as well as types. A valid token is defined as a token which has an accurate phonetic spelling representation as is spoken or pronounced; while an invalid token is a token with a spelling error, typographical error, or a foreign word with varied spellings, abbreviations and so on. We then store those correct tokens into a database as standard or correct words. In our sample there found to be 94,239 tokens and 18,250 types which are considered as valid. However, the invalid tokens are 64,040 and only 1,741 types.

The mechanism of our algorithm as described in Fig.1.4 starts by allowing running text as input. The text is then tokenized and checked in the validation database (database1). If an input token is recognized, the system generates it as an output. If the database doesn't identify a word as valid or error free, then the input word is checked in Database 2 where suggested most frequent invalid words are stored and given their appropriate correct and canonical corresponding. The algorithm classifies these words into regular erred words and irregular erred words. A rule-based method is applied for all regular erred words and those irregular words are normalized using matching-replacement approach. The output is a correct canonical or standardized form that can be used by any NLP application.

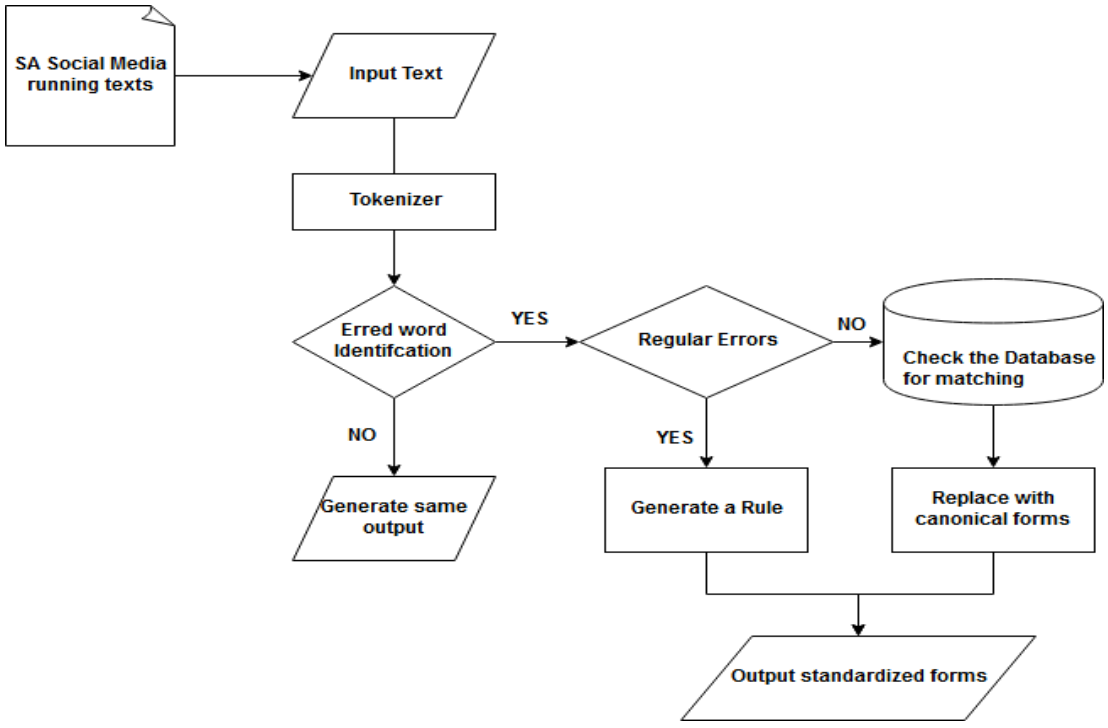


**Fig. 1.4 Shows the mechanism of our algorithm**

## VII. RESULTS AND EVALUATION

The main result of this paper is summarized by developing a corpus normalizer for SA social media data. This normalizer as seen in Fig.1.5 is a user-friendly system consisting of input-output interface. Since there are no previous normalizers

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 7, Issue 2, pp: (428-436), Month: April - June 2019, Available at: www.researchpublish.com

for comparison and evaluation, the system is tested and evaluated on a new corpus of the same dialect consisting of 66,012 tokens and 14,339 types. The system succeeded to normalize 99% of the tokens and 98% among types (i.e., 65,336 tokens and 14,132 types are normalized correctly). However, it fails to tackle 1% among tokens and 2% among the types (i.e., 676 tokens and 207 types). The reason behind this is that the data is less frequent. However, for improvement, we can achieve higher percentage or accuracy by training such data.


**Fig 1.5 Shows a demonstration of our normalizer performance**

## VIII. CONCLUSION

In this paper, we developed our Normalizer for San'ani Arabic Social media texts that are extracted from Facebook and Telegram apps. The normalizer is considered the first in the area of research that addresses such problems. Our normalizer is limited for dealing with San'ani Arabic spoken in Yemen. However, it can be adapted for other dialects. Our classification technique as well as error identification process that we adapt in this study facilitates very much in feeding our system and building our algorithm; and hence, obtaining high accuracy. This system will be an essential tool for preparing the dialectal texts for any further NLP applications.

## REFERENCES

[1] Liu, Fei, Fuliang Weng, and Xiao Jiang. "A broad-coverage normalization system for social media language." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 1035-1044. Association for Computational Linguistics, 2012.

[2] Farzindar, Atefeh, and Diana Inkpen. "Natural language processing for social media." *Synthesis Lectures on Human Language Technologies* 8, no. 2 (2015): 1-166.

[3] Lusetti, Massimo, Tatyana Ruzsics, Anne Göhring, Tanja Samardžić, and Elisabeth Stark. "Encoder-Decoder Methods for Text Normalization." In Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), pp. 18-28. 2018.

[4] Crystal, David. *Txtng: The gr8 db8*. OUP Oxford, 2008.

[5] Habash, Nizar Y. "Introduction to Arabic natural language processing." *Synthesis Lectures on Human Language Technologies* 3, no. 1 (2010): 1-187.

[6] Buckwalter, Timothy. "Issues in Arabic morphological analysis." In *Arabic computational morphology*, pp. 23-41. Springer, Dordrecht, 2007.

[7] Vilar, David, Jan-T. Peter, and Hermann Ney. "Can we translate letters?." In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 33-39. Association for Computational Linguistics, 2007.

[8] Samardzic, Tanja, Yves Scherrer, and Elvira Glaser. "Normalising orthographic and dialectal variants for the automatic processing of Swiss German." (2015).

[9] ERYİĞİT, GÜLŞEN, and D. İ. L. A. R. A. TORUNOĞLU-SELAMET. "Social media text normalization for Turkish." *Natural Language Engineering* 23, no. 6 (2017): 835-875.

[10] Watson, Daniel, Nasser Zalmout, and Nizar Habash. "Utilizing Character and Word Embeddings for Text Normalization with Sequence-to-Sequence Models." *arXiv preprint arXiv:1809.01534* (2018).

[11] Nawar, Michael. "CUFE $@ $ QALB-2015 Shared Task: Arabic Error Correction System." In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pp. 133-137. 2015.