Performance Comparison of Bootstrapped Statistical Taggers on Urdu Tweets

Amber Baig[*], Mutee U Rahman[*], Sehrish Abrejo[*], Khalid H Mohamadani[**], Ahsanullah Baloch[*]

[*] Department of Computer Science, Isra University, Hyderabad, Pakistan
[**] School of Electronics, Beijing University of Posts and Telecommunications, China

DOI: 10.29322/IJSRP.11.07.2021.p11559
http://dx.doi.org/10.29322/IJSRP.11.07.2021.p11559

Abstract- Twitter, a social media platform has experienced substantial growth over the last few years. Thus, huge number of tweets from various communities is available and used for various NLP applications such as Opinion mining, information extraction, sentiment analysis etc. One of the key pre-processing steps in such NLP applications is Part-of-Speech (POS) tagging. POS tagging of Twitter data (also called noisy text) is different than conventional POS tagging due to informal nature and presence of Twitter specific elements. Resources for POS tagging of tweet specific data are mostly available for English. Though, availability of tagset and language independent statistical taggers do provide opportunity for resource-poor languages such as Urdu to expand coverage of NLP tools to this new domain of POS tagging for which little effort has been reported. The aim of this study is twofold. First, is to investigate how well the statistical taggers developed for POS tagging of structured text fare in the domain of tweet POS tagging. Secondly, how can these taggers be used to overcome the bottleneck of manually annotated corpus for this new domain. To this end, Stanford and MorphoDiTa taggers were trained on 500 Urdu tweet gold-standard corpus and were utilized for semi-automatic corpus annotation in bootstrapped fashion. Five bootstrapping iterations for both the taggers were performed. At the end of each iteration, the performance of taggers was evaluated against the development set and automatically tagged, manually corrected 100 tweets were added in the training set to retrain both models. Finally, at the end of last iteration, tagger performance was evaluated against test set. Stanford tagger achieved an accuracy of 93.8% Precision, 92.9% Recall and 93.3% F-Measure. Whereas, MorphoDiTa tagger achieved an accuracy of 93.5% Precision, 92.6% Recall and 93% F-Measure. A thorough error analysis on the output of both taggers is also presented.

Index Terms- Bootstrapping, manual annotation, natural language processing, POS tagging, statistical taggers, Urdu tweets.

I. Introduction

Twitter, a popular microblogging platform has grown substantially in recent years. With more than 326 million users, Twitter permits its users to post limited sized tweets having maximum 280-characters about wide variety of topics. These tweets can be exploited further for a range of activities like forecasting or explaining real-world outcomes through opinion mining, targeted advertisement campaigns through mining users' interests, getting client feedback on brands, government regulations, and other topics [3].

POS tagging which marks the words of a text with syntactic information such as noun, pronoun, verb, etc. [4], is an important prerequisite for above listed and many more Natural Language Processing (NLP) applications [2]. However, most POS taggers are developed for tagging standardized, grammatically edited texts such as newspaper articles and literary writings. Tweets, however, are different from such standard text due to their informal writing style. Because of 280-character limit, tweets contain various forms of non-standard language like use of special elements such as emoticons and hashtags, lengthening and shortening of words, and phrase abbreviation [5]. In addition to these explicit attributes, tweets may also have unusual grammar patterns with unintentional spelling mistakes [5-6], [13].

Recent availability of tagset and language independent statistical POS taggers made it easier to expand the coverage of POS tagging to new domains and languages. These taggers are robust and produce state-of-the-art results. Nevertheless, for training and performance testing, these statistical taggers rely on large amount of manually annotated gold standard corpora which is not always available for several languages and domains. Manual annotation is costly and takes time [19]. The question of annotation costs and ways to minimize the dependence on such annotated corpora has been a recurring theme in the NLP community for the last two decades [7], [10]. Researchers require reliable strategies to accelerate the annotation process while avoiding biasing the obtained gold standard [8].

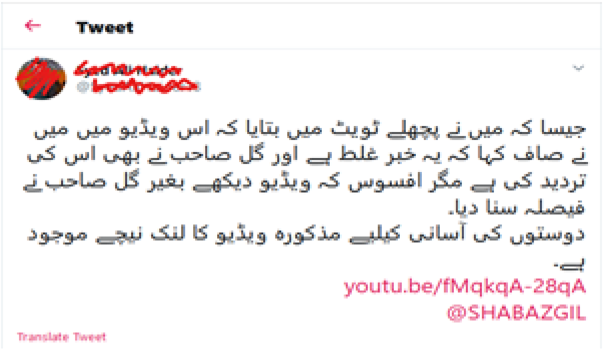
Semi-automatic corpus annotation is one such method where a trained PoS tagger is used to label raw text, which is then hand-corrected by human annotators [9]. Thus, development of large POS tagged corpus can be achieved in short time. Besides, statistical POS taggers assist to bootstrap the annotation process, since their performance will improve as the size of the training corpus increases. Since performance of NLP applications is greatly influenced by the nature of processed text [2], it is an open question how well NLP tools and techniques used for structured newswire data will perform for Twitter for understanding and exploiting tweets; since POS tagging of tweets differs greatly from that of more formal texts [1]. Therefore, this paper focuses on experimental evaluation of the performance of two state-of-the-art statistical POS taggers for Urdu tweets. Urdu is national language of Pakistan and is a significant language of South Asia with a large speaker base of more than 300 million worldwide [12]. Due to the experiments on numerous NLP tasks, Urdu Language Processing has recently been the current research trend. Despite these attempts, Urdu continues to be a low-resourced language, specially POS tagging of Urdu tweets is still in its early stages.

This paper builds upon the work presented in [3] and evaluates how Stanford tagger and MorphoDiTa tagger, both language independent statistical taggers, can be utilized to create POS tagged Urdu tweet corpus using bootstrapped semi-supervised approach and compares both taggers' performance for Urdu tweets POS tagging. The paper is structured as follows: Section 2 enlists characteristics of Urdu tweets. In Section 3, materials and methods of this study are described. Results and their discussion are outlined in Section 4 and lastly, conclusions are presented in Section 5.

II. CHARACTERISTICS OF URDU TWEETS

Following [5-6], [13], based on intentionality or communication requirements motivating word variants, peculiarities of Urdu tweets were analyzed and found to have properties which are summarized below:

Along with standard words (e.g. name of a place, person, object, or building, etc.), Urdu tweets were found to have Twitter-specific elements such as mention, reply, hashtag, retweet and url. A sample tweet with url and mention is shown in Fig. 1.

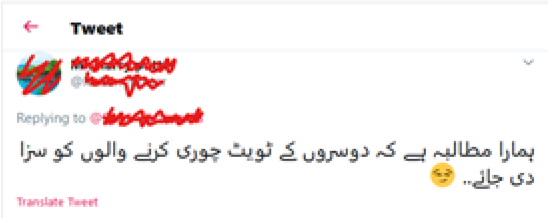


[redacted]
[redacted]

جیسا کہ میں نے پچھلے ٹویٹ میں بتایا کہ اس ویڈیو میں میں
نے صاف کہا کہ یہ خبر غلط ہے اور گل صاحب نے بھی اس کی
تردید کی ہے مگر افسوس کہ ویڈیو دیکھے بغیر گل صاحب نے
فیصلہ سنا دیا۔
دوستوں کی آسانی کیلیے مذکورہ ویڈیو کا لنک نیچے موجود
ہے۔
youtu.be/fMqkqA-28qA
@SHABAZGIL

Figure 1. Example of Tweet containing URL and Mention

Use of emoticons and emojis is also found common in Urdu tweets for expressing the emotions or feelings. Fig. 2 shows a tweet with an emoticon.



[redacted]
[redacted]
Replying to @[redacted]

ہمارا مطالبہ ہے کہ دوسروں کے ٹویٹ چوری کرنے والوں کو سزا
دی جائے.. 😌

Figure 2. Example of Tweet containing an emoticon

Concatenated words where space between separate words is omitted. Thus, combining these words into a single token were also observed to be used quite frequently in Urdu tweets. Users may intentionally or accidentally connect words to overcome the tweets' limited length. For example, 'کرناہوگا' instead of 'کرنا ہو گا' and 'کیامقصدہے' instead of 'کیا مقصد ہے', etc.

III. METHODS AND TOOLS

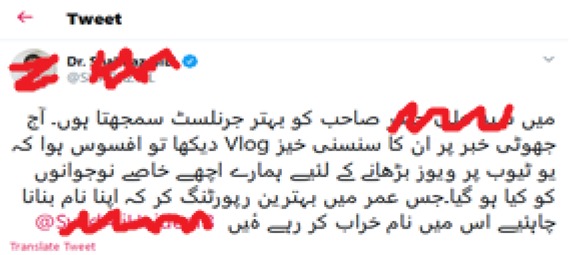
Likewise, unnecessary space insertion between tokens of a single word breaking it into two separate words (e.g. 'حال آنکہ' instead of 'حالانکہ', 'نا اہل' instead of 'نااہل', etc) was also observed. Users may segment words deliberately or accidentally. Repeated Letters i-e letters intentionally repeated for expressing subjectivity and emotion by users were also recurring in tweets. Examples include, 'اووووہ', 'ارررررے', '-------- اچھا', 'کیا ؟؟؟؟؟؟', etc.

Another recurring property of Urdu tweets is of spelling mistakes and spelling variation. It is not easy to know user's intention, but some words may accidentally have been misspelled. Similarly, different users may write same words using different spellings. For example, 'لیکر' instead of 'لے کر', 'بعدماشی' instead of 'بدمعاشی', 'کوشیش' instead of 'کوشش', 'چھڑنے' instead of 'چھیڑنے', etc.

Informal slangs which are usually limited to a specific context or group of people were also present in Urdu tweets. For example, 'ٹوپی ڈرامہ' or 'چونا لگانا' for fooling someone, 'اسکرو ڈھیلا' for stupid or insane person, 'جگاڑ' for finding easy solution, 'لش' for wonderful, etc.

The rich morphological nature of Urdu allows its users to borrow some words and multi words abbreviations from English by using their Urdu transliteration in tweets. For example, 'ٹرول/Troll', 'موسٹ/Most', 'انیشیٹو/Initiative', 'ایٹی ٹیوڈ/Atitude', etc.

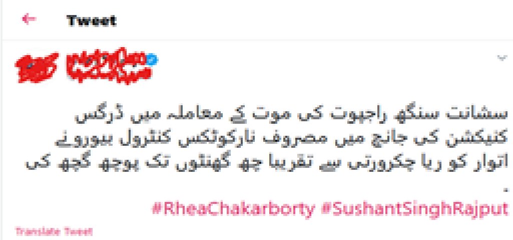
Some Urdu tweets having foreign words from local or international languages, especially English. Were also observed. Locations, events, English hashtags or retweeted English tweets may be indicated by these words along with Urdu comments as shown in Fig 3 and Fig 4.



Dr. [redacted]
@[redacted]

میں [redacted] صاحب کو بہتر جرنلسٹ سمجھتا ہوں. آج
جھوٹی خبر پر ان کا سنسنی خیز Vlog دیکھا تو افسوس ہوا کہ
یو ٹیوب پر ویوز بڑھانے کے لئیے ہمارے اچھے خاصے نوجوانوں
کو کیا ہو گیا.جس عمر میں بہترین رپورٹنگ کر کہ اپنا نام بنانا
چاہئیے اس میں نام خراب کر رہے ہٰیں @[redacted]

Figure 3. Example of Tweet containing English Words



[redacted]

سشانت سنگھ راجپوت کی موت کے معاملہ میں ڈرگس
کنیکشن کی جانچ میں مصروف نارکوٹکس کنٹرول بیورونے
اتوار کو ریا چکرورتی سے تقریبا چھ گھنٹوں تک پوچھ گچھ کی
.
#RheaChakarborty #SushantSinghRajput

Figure 4. Example of Tweet containing English Hashtags

http://dx.doi.org/10.29322/IJSRP.11.07.2021.p11559

A. POS Tagset and Corpus

UNTPOS (Urdu Noisy Text Part of Speech), a POS tagset consisting of 33 tags designed specifically for POS tagging of typical parts of speech (nouns, adjectives, etc.) as well as collection of token discrepancies primarily found in Urdu tweets is used in this study. For full review of tagset, refer to [3]. The authors also publicly made available their manually annotated gold standard POS tagged corpus of 500 tweets which can be used for training of any statistical tagger along with pre-processed raw corpus of 4500 tweets on various topics, for further experiments on Urdu tweets.

For this study, 500 tweets from the pre-processed raw tweet corpus were randomly selected for bootstrapping tagger experiments and the gold standard corpus is used for initial tagger training.

B. Statistical Taggers

Since the objective of this study is to evaluate performance of statistical POS taggers on Urdu tweets, two state-of-the-art taggers were selected. The choice of the taggers is based on their language and tagset independence. The first tagger is the well-known Stanford tagger [15] which is written in Java and creates statistically trained model based on tags in a training dataset.

MorphoDita (Morphological Dictionary and Tagger) open source tagger [14] is the second tagger. The tagger is made up of several modules, including a tokenizer, a morphological generator, and a morphological analysis and morphological tagger module [11], it also makes use of an accessible and up-to-date morphological dictionary called MorfFlexCZ [18].

C. Evaluation Metrics

For evaluating tagger performance, conventional metrics of precision, recall and f-measure are used. Their calculation is described in the following equations:

$$\text{Precision(P)} = \frac{\text{Correctly tagged tokens}}{\text{Total tagged tokens}} \qquad (1)$$

$$\text{Recall(R)} = \frac{\text{Correctly tagged tokens}}{\text{Total possible correctly tagged tokens}} \qquad (2)$$

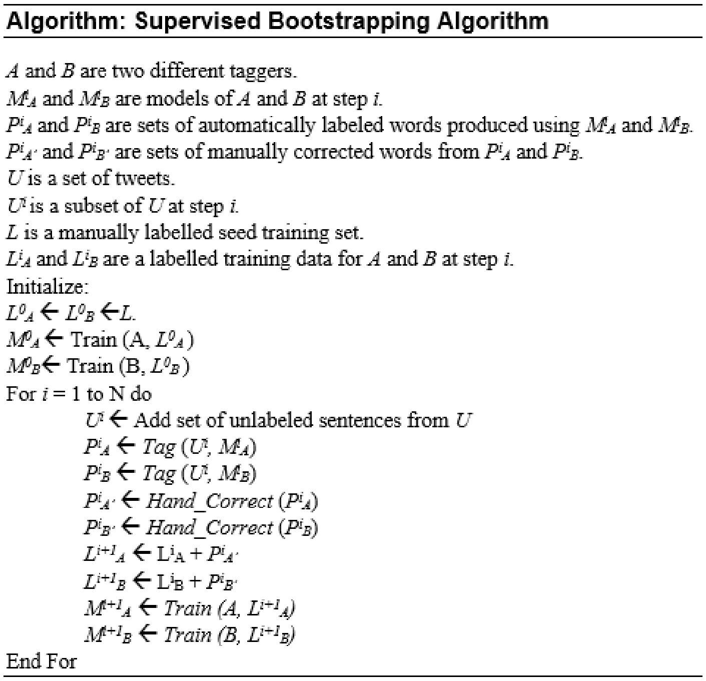
$$\text{F-Measure} = 2 \times \frac{\text{P}\times\text{R}}{\text{P+R}} \qquad (3)$$

D. Experiments

Bootstrapping is the process of creating annotated training data from unannotated large amounts of data [16]. In this approach, a pretrained tagger is used to tag the words of the sentences which may then be manually corrected and included in the gold standard training corpus. The tagger is then retrained using this corpus to label next batch of sentences and this process of retraining and correction continues iteratively until some termination criteria is met [17].

Since this work builds upon the work of [3], same 500 pre-processed tweets (13,643 tokens) used in [3] were selected for performing bootstrapping experiments. 500 tweet gold standard corpus was segmented into 300 tweets training set, and 100 tweets development set and test set respectively. Two separate POS tagger models were trained using Stanford POS tagger and MorphoDiTa tagger using 300 sentence training data. These models were evaluated against development set for their performance accuracy and to establish baseline scores for both the taggers against which future models' performance be compared. The resultant baseline scores are shown in Table 1.

Next, five bootstrapping iterations were carried out with the help of both models to label 100 tweets in every iteration. The algorithm used for bootstrapping POS tagger training is listed in Fig. 5.



```
Algorithm: Supervised Bootstrapping Algorithm

$A$ and $B$ are two different taggers.
$M^i_A$ and $M^i_B$ are models of $A$ and $B$ at step $i$.
$P^i_A$ and $P^i_B$ are sets of automatically labeled words produced using $M^i_A$ and $M^i_B$.
$P^i_{A'}$ and $P^i_{B'}$ are sets of manually corrected words from $P^i_A$ and $P^i_B$.
$U$ is a set of tweets.
$U^i$ is a subset of $U$ at step $i$.
$L$ is a manually labelled seed training set.
$L^i_A$ and $L^i_B$ are a labelled training data for $A$ and $B$ at step $i$.
Initialize:
$L^0_A \leftarrow L^0_B \leftarrow L$.
$M^0_A \leftarrow$ Train (A, $L^0_A$)
$M^0_B \leftarrow$ Train (B, $L^0_B$)
For $i$ = 1 to N do
        $U^i \leftarrow$ Add set of unlabeled sentences from $U$
        $P^i_A \leftarrow Tag(U^i, M^i_A)$
        $P^i_B \leftarrow Tag(U^i, M^i_B)$
        $P^i_{A'} \leftarrow Hand\_Correct(P^i_A)$
        $P^i_{B'} \leftarrow Hand\_Correct(P^i_B)$
        $L^{i+1}_A \leftarrow L^i_A + P^i_{A'}$
        $L^{i+1}_B \leftarrow L^i_B + P^i_{B'}$
        $M^{i+1}_A \leftarrow Train(A, L^{i+1}_A)$
        $M^{i+1}_B \leftarrow Train(B, L^{i+1}_B)$
End For
```

Figure 5. Bootstrapping Algorithm

At the end of every cycle, the same annotators who annotated the gold standard Urdu tweet corpus of [3], hand corrected these labelled tweets. These rectified tweets were then added to the training set, which was used for retraining new tagger models. After that, the freshly trained models were used to tag the next 100 tweets, and their accuracy was compared on the development set. The performance of both taggers is evaluated against the test set at the end of the fifth and final iteration.

IV. RESULTS AND DISCUSSION

Result of bootstrapping experiments are shown in Table 1. The highest scoring models occur on the fifth repetition, with Stanford Tagger scoring 92.5 percent precision, 93.5 percent recall, and 93 percent f-measure and MorphoDiTa Tagger scoring 92.3 percent precision, 91.4 percent recall, and 91 percent f-measure. Stanford outperformed MorphoDita but the difference in accuracy score is very little. All results are calculated on the development set.

On the test set of 100 tweets (2,306 tokens), overall error percentage was 12.4 percent for Stanford and 12.9 percent for MorphoDiTa, respectively. Error analysis shows that low-frequency words, ambiguous words, and unseen words are the three main categories of POS tagging errors. A comparison of error analysis for both taggers is shown in Figure 6.

http://dx.doi.org/10.29322/IJSRP.11.07.2021.p11559

Table I. Bootstrapping Parser Evaluation

Taggers	Evaluation Metric	Baseline	Iterations					Final Evaluation
			1	2	3	4	5	
Stanford	Precision	84.3	86.6	87.9	89	90.3	92.5	93.8
	Recall	80.4	88.6	89.8	90	92.3	93.5	92.9
	F-Measure	82.3	87.6	88.8	89.5	91.3	93	93.3
MorphoDiTa	Precision	83.7	84.9	85.5	88.7	91.1	92.3	93.5
	Recall	81.8	85.2	88.5	87.7	89.1	91.4	92.6
	F-Measure	82.7	84.8	87	88.2	90.1	91.8	93

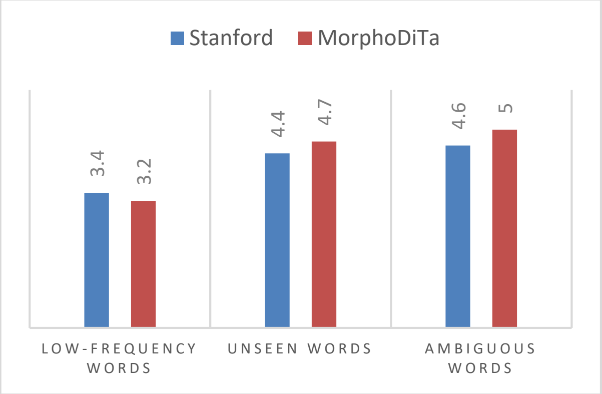


Figure 6. Stanford & MorphoDiTa Taggers' Error Analysis

Low-frequency words are known to be more problematic for learning and correct prediction. Low-frequency terms, such as "ٹرول" (troll), "اوسم" (awesome), etc., as well as slangs, links, and hashtags, were found to be a source of tagging mistakes in the test set. There was also a pattern of emoticon misclassification and, in certain situations, punctuation. It has also been observed that MorphoDiTa tagger had difficulty in tagging Twitter usernames starting with @ even though training set had labelled examples of Twitter usernames. Sometimes these were correctly labelled as proper noun "PROPN", whereas most of the instances, these usernames were incorrectly tagged as "X".

Stanford tagger on the other hand, always labelled them correctly. In the test set, these errors were mostly caused by emoticons and punctuation marks sequences (e.g., 😊😍😉😘😜😚, !!!!). Whereas in training data, these sequences were not present. Low-frequency words' error rate in case of Stanford is 3.4% whereas for MorphoDita, it is 3.2%.

Words absent in training set but appear in the test set are known as unseen words. Most of these unseen words in the test set are emoticons, named entities, Urdu transliterated English words, and slangs among other elements. Standford has a 4.4 percent error rate for unknown words, while MorphoDiTa has a 4.7 percent error rate.

The major cause of tagging errors in the test set was found to be the ambiguous words both for Standford and MorphoDiTa taggers with 4.6% and 5% respectively. Numerous explanations can be associated for the occurrence of ambiguous words. First, Urdu Tweets are mostly conversations and some words partake different forms while writing. For example, "امریکہ vs امریکا", "ن لیگ vs نون لیگ", "سعودیہ vs عرب سعودی" etc. Likewise, unnecessary space insertion also causes tagging mistakes. For example, the words "نااہل" and "بےضرر" are basically adjectives but labelled incorrectly as "PART" and "NOUN" due to space inserted between. Another source of tagger error is where two or more words are merged, like "اس لیے" written as "اسلیے" resulting in PRON and ADP being incorrectly labelled PRON. Same thing happened when the words "اس طرح" misspelt as "اسطرح", thus classified as "NOUN" rather than a "DET" and NOUN. Handling of punctuation, spelling mistakes and special characters was also problematic in the same way. One frequently occurred case of spelling mistake was SCONJ "کہ" written as "کے" in the test set. Thereby, labelled incorrectly as "ADP" by taggers. These error patterns are found in both the taggers being compared.

One major problem that taggers face is confusion between words having similar word form, but different meaning based on use. One such case is amongst subordinate conjunction "تو", pronoun "تو" and particle "تو". Both taggers labelled PART "تو" and PRONP "تو" as "SCONJ" in majority of situations. Another example is the use of certain determinative articles and pronouns. "یہ" is DET in "یہ شاگرد فرانسیسی زبان بولتا ہے / This student speaks French", where as in "یہ فرانسیسی زبان بولتا ہے / He speaks French", "یہ" is a pronoun. Again, both taggers labelled "یہ" as DET regardless of the context. Similar case was observed for DET "یا" in "یا اللہ" and CCONJ "یا".

Another recurrent mistake found in tagging is misinterpretation between adjectives (ADJ), nouns (NOUN) and proper nouns (PROP). This is common tagging error as many nouns can interchangeably be used as both proper and common nouns. In Urdu, "حنا", "اظہار", "ڈار" etc., can be treated as both nouns and proper nouns depending on the context of use.

Table 1 shows that bootstrapping POS taggers is effective in general. In comparison to manually annotating data, auto-tagging and manual rectification take less time, while finally induced models attained good results.

V. CONCLUSION

Language and tagset independent statistical POS taggers provide opportunity for resource-poor languages like Urdu to expand the domain of their NLP applications to new areas. Twitter POS tagging is one such area which is still in its early stages of research and development for Urdu. Due to which, large amount of manually annotated corpus for statistical taggers is deficient. In this paper, semi-automatic corpus annotation is investigated as a potential means to overcome this lack of annotated corpora for Urdu tweets. Stanford and MorphoDiTa, two state-of-the-art statistical POS taggers were used to bootstrap the annotation process and their performance have been compared for POS tagging of Urdu tweets. Both the taggers were evaluated using same training, development, and test data. At the end of experiment, 93.8% Precision, 92.9% Recall and 93.3% F-Measure in terms of accuracy was scored by Stanford tagger. Whereas, MorphoDiTa tagger achieved an accuracy of 93.5% Precision, 92.6% Recall and 93% F-Measure. The error analysis performed on both taggers reveled that Stanford tagger outperformed MorphoDiTa in tagging unseen words and ambiguous words.

Though, Stanford taggers' performance was slightly better than MorphoDiTa, the results reported in Table 1 show that there is not much performance difference between both the taggers. The results reported in Table 1 indicated that this method is effective for increasing the amount of training corpora in less time as compared to manually annotating the corpus from scratch. For future work, other statistical taggers can also be used along with exploring semi-supervised and unsupervised methods for tagger training and corpus annotation.

ACKNOWLEDGMENT

Isra University, Hyderabad, Pakistan partially supports this work. All faculty members of Department of Computer Science, Isra University are acknowledged for their help and support throughout the course of this study.

REFERENCES

[1] F. Albogamy and A. Ramsay, "POS tagging for Arabic tweets," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 2015, pp. 1-8.

[2] W. AlKhwiter and N. Al-Twairesh, "Part-of-speech tagging for Arabic tweets using CRF and Bi-LSTM," *Computer Speech & Language,* vol. 65, p. 101138, 2021.

[3] A. Baig, M. U. Rahman, H. Kazi, and A. Baloch, "Developing a POS Tagged Corpus of Urdu Tweets," *Computers,* vol. 9, p. 90, 2020.

[4] D. Briesch, R. Hobbs, C. Jaja, B. Kjersten, and C. Voss, "Training and evaluating a statistical part of speech tagger for natural language applications using kepler workflows," *Procedia Computer Science,* vol. 9, pp. 1588-1594, 2012.

[5] J. Eisenstein, "What to do about bad language on the internet," in *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, 2013, pp. 359-369.

[6] J. Foster, "“cba to check the spelling”: investigating parser performance on discussion forum posts," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 381-384.

[7] D. Garrette and J. Baldridge, "Learning a part-of-speech tagger from two hours of annotation," in *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2013, pp. 138-147.

[8] T. Lingren, L. Deleger, K. Molnar, H. Zhai, J. Meinzen-Derr, M. Kaiser, *et al.*, "Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements," *Journal of the American Medical Informatics Association,* vol. 21, pp. 406-413, 2014.

[9] H. Loftsson, "Correcting a PoS-tagged corpus using three complementary methods," in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 2009, pp. 523-531.

[10] G. Ngai and D. Yarowsky, "Rule writing or annotation: Cost-efficient resource usage for base noun phrase chunking," presented at the The 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, 2001.

[11] P. Pořízka, *Tvorba korpusů a vytěžování jazykových dat: metody, modely, nástroje*: Vydavatelství Filozofické fakulty Univerzity Palackého v Olomouci, 2014.

[12] A. A. Raza, A. Habib, J. Ashraf, and M. Javed, "A review on Urdu language parsing," *Int. J. Adv. Comput. Sci. Appl,* vol. 8, pp. 93-97, 2017.

[13] D. Seddah, B. Sagot, M. Candito, V. Mouilleron, and V. Combet, "The French Social Media Bank: a treebank of noisy user generated content," in *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, 2012, pp. 2441–2458.

[14] J. Straková, M. Straka, and J. Hajic, "Open-source tools for morphology, lemmatization, POS tagging and named entity recognition," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 13-18.

[15] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 252-259.

[16] F. Albogamy, A. Ramsay, and H. Ahmed, "Arabic tweets treebanking and parsing: A bootstrapping approach," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, 2017, pp. 94-99.

[17] F. Zarei, A. Basirat, H. Faili, and M. Mirain, "A bootstrapping method for development of Treebank," *Journal of Experimental & Theoretical Artificial Intelligence,* vol. 29, pp. 19-42, 2017.

[18] J. Hajič and J. Hlaváčová, "MorfFlex CZ," 2013.

[19] C. Callison-Burch, "Fast, cheap, and creative: Evaluating translation quality using Amazon’s Mechanical Turk," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 286-295.

AUTHORS

First Author – Amber Baig, PhD (Computer Science), Isra University, Pakistan. amberbaig@gmail.com
Second Author – Mutee U Rahman, PhD (Computer Science), Isra University, Pakistan. muteeurahman@gmail.com
Third Author – Sehrish Abrejo, M.Phil (Computer Science), Isra University, Pakistan. sehrish-abrejo@hotmail.com
Fourth Author – Khaild H Mohamadani, PhD Scholar, School of Electronics, Beijing University of Posts and Telecommunications, China. wangkhm@bupt.edu.cn
Fifth Author – Ahsanullah Baloch, PhD (Computational Mathematics), Isra University, Pakistan. ahsanullah.baloch@isra.edu.pk

Correspondence Author – Amber Baig, amberbaig@gmail.com, amber.baig@isra.edu.pk

http://dx.doi.org/10.29322/IJSRP.11.07.2021.p11559