



## RESEARCH PAPER

### Investigating the Variation of Formal and Informal Vocabulary in Pakistani English on Twitter

<sup>1</sup>Khubaib ur Rehman, <sup>2</sup>Dr. Tayyaba Yasmin and <sup>3</sup>Sidra Bukhari

1. Research Associate, Riphah Institute of Language and Literature, Riphah International University Gulberg Campus Lahore, Punjab, Pakistan
2. Associate Professor, Department of English, University of Education, Lahore, Punjab, Pakistan.
3. Lecturer, Department of English, University of Education, Lahore, Punjab, Pakistan.

**\*Corresponding Author** | khubaib.rehman@riphah.edu.pk

## ABSTRACT

This study investigates the vocabulary of the English language and its variation of formality and informality within a specified sample of tweets, focusing on the social modulation of linguistic forms in digital communication. Using a mixed-method approach, the study analyzes Twitter data through MAT tagger and corpus analysis tools. A pilot test was conducted to validate encoding using UTF-16 in Python due to Urdu script interference. Quantitative analysis employed a differential scale (20\*-20) on both axes, while qualitative interpretation was based on Halliday's (1960) Register Theory. Findings reveal that formality and informality vary situationally, shaped by social and linguistic constraints. The quantitative outcomes show distinguishable lexical shifts between explicit and situationally dependent expressions, indicating social media's influence on language construction. The results also confirm that bilingual interference moderates linguistic variability within digital discourse. Future research should expand corpus diversity across platforms and demographics to refine linguistic dimensionality.

**KEYWORDS** Linguistic Variation, Vocabulary, Dimensional Analysis, Twitter Corpus, Register Theory

## Introduction

Language and communication have endured dramatic change due to the development of social media, which has led to the development of new language forms that include digital, computer-mediated language. One of the largest social networking platforms is Twitter, which is more popular among the masses to share their experiences, feelings, and thoughts on different aspects of life through text to examine the use of formal and informal vocabulary and language variation in any social group.

The objective of the current paper is to examine the variation of formal and informal vocabulary in Pakistani English-based tweets. To analyze and consider the formal and informal vocabulary in Pakistani English tweets, the corpus methodology is being employed. The article considered the contextual factors that make it effective to use formal and informal vocabulary in Twitter posts.

The significance of this study lies in the exploration of linguistic variation in the PE Twitter corpus that has not been properly analyzed before. Moreover, the research is conducive to determining the language use and pervasiveness of English language learning and teaching practices in a Pakistani context through social media.

Objectives that are associated with data streamline the social construction of language variation and a thematic sketch of linguistic construction manipulated with a biased resistance. Register theory suggests that speakers use language differently in different situations. It implies that language use is affected by situational factors composed of circumstances, aims and audience of communication, and individual factors which include age, level of education, gender and socioeconomic position.

A wide range of both formal and informal vocabulary in Pakistani English on Twitter can be examined through register theory, which can present a structure for examining how various registers of language are utilized on a system. The patterns of language used in formal or informal registers can be discovered by assessing the linguistic characteristics and stylistic selections of tweets.

Formal language is often categorized by its complicated syntax, professional terminologies, and standard rules of grammar, in comparison to informal language which is commonly defined by the use of colloquial language, slang, and atypical language. Register theory furthermore claims that the language use is adaptable varying according to circumstances and target audience. The purpose of the tweets, its targeted audience and the user's social identity are the factors through which changes in the use of formal and informal language on Twitter can be detected.

Differences in the formal and informal language in Pakistani English on Twitter can be analyzed and can highlight the ever-changing characteristic of language use in online conversations, through the productive framework of register theory.

## **Literature Review**

A thesis of PhD was conducted by Maybaum (2015) at the University of Hafsa on Twitter's community-specific lexical inventions that were examined in a corpus of 19 million words. The analysis showed that the creative variations spread over time in an S-shaped pattern. This pattern resembles the diffusion of social innovations more generally, indicating that some degrees of linguistic change can be seen as falling within the scope of general human social behaviour.

As a result, sociological models of innovation diffusion can be applied to specific cases of linguistic diffusion. Systematic diversity in the diffusion patterns shows that the social process of innovation adoption takes place over time in several discrete phases as opposed to occurring instantly (Ali, 2009). Even while the primary meanings and purposes of the Twitter People variable as a whole were cooperatively negotiated, distinct Twitter People variations' conversation patterns were disproportionately influenced by individual Twitter users. Users of Twitter adopted creative linguistic variations to create communal frameworks for time and space as well as social conventions. To offer a model of language change as an ongoing social process, theoretical components from innovation diffusion and sociolinguistics research traditions are merged. Individual speakers appropriate underspecified linguistic variations for the satisfaction of unmet social demands. As a result of both individual speaker agency and cooperative negotiation (through the stages of the innovation-decision process), which results in a collective understanding of social meaning, linguistic change is triggered (Lukin et.al, 2011).

Pak and Paroubek (2010) elaborated on the contribution of microblogging which has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life every day. Therefore, microblogging websites are rich sources of data for opinion mining and sentiment analysis. Because microblogging has

appeared relatively recently, there are a few research works that were devoted to this topic (Aziz & Ali, 2018). In their studies they focus on using Twitter, the most popular microblogging platform, for the task of sentiment analysis to show, how to automatically collect a corpus for sentiment analysis and opinion mining purposes.

By performing linguistic analysis of the collected corpus and explaining the discovered phenomena, this study generates a sentiment classifier that can determine positive, negative and neutral sentiments for a document. Experimental evaluations show that the proposed techniques are efficient and perform better than previously proposed methods (Khan & Ahmad, 2020). In the research, Pak and Paroubek worked with English, however, the proposed technique can be used with any other language.

Experimental testing demonstrates that the suggested procedures are effective and outperform previously suggested approaches (Rana & Tariq). Although the employed language is English in Pak and Paroubek research, the suggested approach may be applied to any other language.

Mahmood (2009) established his PhD thesis on Pakistani English from Bahauddin Zakariya University Multan which has spread over the world and almost every country now uses it for domestic and international communication. It is assisting in bridging language and ethnic divides around the world. Despite its worldwide standing, it nonetheless displays elements of local culture. It is widely used, and the notion that it primarily represents Judo-Christian cultural traditions has long been disproved, at least by linguists throughout the world. New dialects and variations are developing as a result of this evolutionary process ( Muhammad Asim Mahmood Supervisor Mubina Talaat, 2009). Around the world, a multitude of research projects are being carried out in the spirit of "World Englishes." Studies are being done on topics including the status of the recently developed variations of English and the parallels and differences between native and nativized versions. In Pakistan, English is regarded as a special official language (Maybaum, 2015).

For both internal and external purposes, English is utilized in the nation. In Pakistan, the English language has been nativized. Insofar as it resembles British and American English, Pakistani English does not only do it not obstruct communication, but it has also acquired its flavor and color.

Pakistani English is not a thoroughly studied dialect. examination of Pakistani English and identify some distinctive traits that set it apart from British and American versions. Essentially, it is a quantitative, corpus-based comparison study. The distinctive characteristics of Pakistani English have been outlined using two million-word corpora from written English in Pakistan, Britain, and America. This study did not fully address the research question, "Is Pakistani English a separate variety?" since there was insufficient evidence to draw any firm conclusions.

Due to its differences from Standard British English, Pakistani English is a non-native variation of the English language. Because language is culture and culture is language, every non-native language has a significant influence on culture (Shaheen & Farooq, 2018). Numerous terms from regional tongues are incorporated into the many varieties of English that are spoken in a given area. Due to Urdu's status as Pakistan's official national language, established by the Islamic Republic of Pakistan's 1973 constitution, cultural fusion in Pakistani English is highly strong. This essay aims to emphasize Urdu terms' incorporation as a distinctive component of Pakistani English (Buckingham, 2008).

Yasir (2009) was of the view that in a sample of literature, including newspapers, official correspondence, and course books chosen using the convenience sampling technique, a qualitative approach has been employed to identify Urduization. The word production process in Pakistani English is highlighted in the research under the umbrella of the Urduization phenomena and how it opens the door for developing code-switched words as a vital component of Pakistani English.

The study also defines neologisms concerning the Pakistani Urduized English that is widely spoken. For this purpose, the eminent works of M. Taalat, Tariq Rehman, Ahmar Mahboob and Baumgardner in the realm of Pakistani English are discussed profoundly. Findings show the Urdu language as a cultural integrant syndicated with Pakistani English explicitly. His study opens the avenues of conspicuous research regarding Pakistani English as a distinct variety of English at the verge of standardization (Abbas, 2018).

### **Material and Methods**

Data is collected from taking samples of twitter posts and all comments are also added in the data till date. As far as the ethical considerations are concerned, the data is based on the political platform which also contains hate speech and abusive language a bit which is an uncontrollable factor. For the concerned dimension analysis, Mat tagger is utilized for the data analysis and encoded from UTF-8 to UTF-16 due to Unicode conversion of Urdu data expressions and for getting more refined results.

The whole encoding process is done with the specified program written for this study on Python for the most possible results outcomes. To narrow down the research, only third dimension is compared with the reference corpus statistics given by Douglas Biber in his book "Variation across speech and writing" published by Cambridge University Press in 1988. In this milieu, the differential statistics are generated for the mean value comparison. Using mix method approach, statistics are quantitatively analyzed with the qualitative description of the deviation is statistics.

### **Results and Discussion**

In the sequence of data analysis, graphs that are based on the Mat tagger are elaborated with qualitative justification of the data with the reference of register variation theoretical approach, introduced by Halliday.

Formal and Informal vocabulary variation under the MAT tagger analysis justify the data more closely with third dimension, explicit vs situation dependent reference having closest genre of political and academic prose. The selected dimension from the all 6 dimension is only to stay close according to the necessities of analysis and to elaborate the one linguistic dimension in more specific way. Figure 1 show the mean of the 8 genres that are scaled on the x-axis with continual interval that remains constant in any data condition

Formal language is often categorized by its complicated syntax, professional terminologies, and standard rules of grammar, in comparison to informal language which is commonly defined by the use of colloquial language, slang, and atypical language. Register theory furthermore claims that the language use is adaptable varying according to circumstances and target audience. The purpose of the tweets, its targeted audience and the user's social identity are the factors through which changes in the use of formal and informal language on Twitter can be detected. Political scenario lower down the bow towards closer use of formal vocabulary while using English language which is prominent

in the results as Halliday suggested the genre relation of language formality and informality (Halliday, 1960).

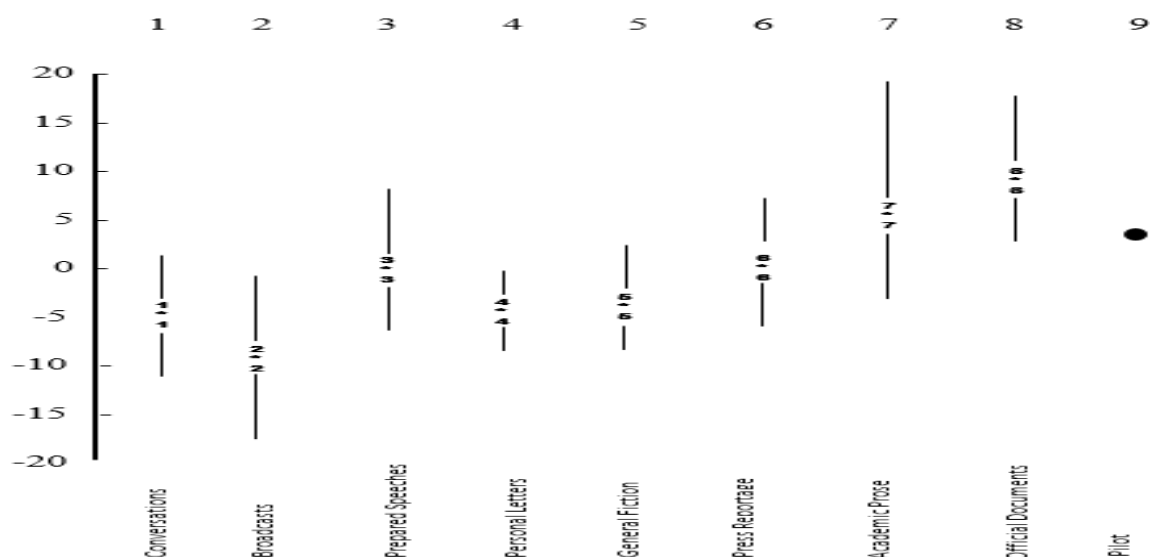


Figure 1: Mean Value Referential

On y-axis the variation the mean is displayed with the (\*) sign to identify the mean value of specified genre with relation to x-axis. Maximum and minimum value of each genre is pointed with starting and ending point of black lines drawn on +y-axis and -y-axis. In the figure 2., the mean lines are drawn from x-axis to y-axis for the cross mean value plotting for the calculation of the statical results. As the 1st genre is conversation, the mean statistic shows the -5 deviation on the differential scale because twitter is not a conversation platform rather than an informational social site. Hence the 2nd genre is broadcast which also violate functionality of twitter as a social platform.

At the 3rd level, the mean of prepared speech project the corrosive result on positive y-axis slightly. The reason is, on the twitter while posting a short paragraph, writer is more conscious about its target either to support or refuse any stance from his or her point of view. On 4th conditioning results of MAT tagger, the mean varies between 0 and -5 because on twitter, there is no consideration of personal letters rather than argument which are always biased with any particular political party.

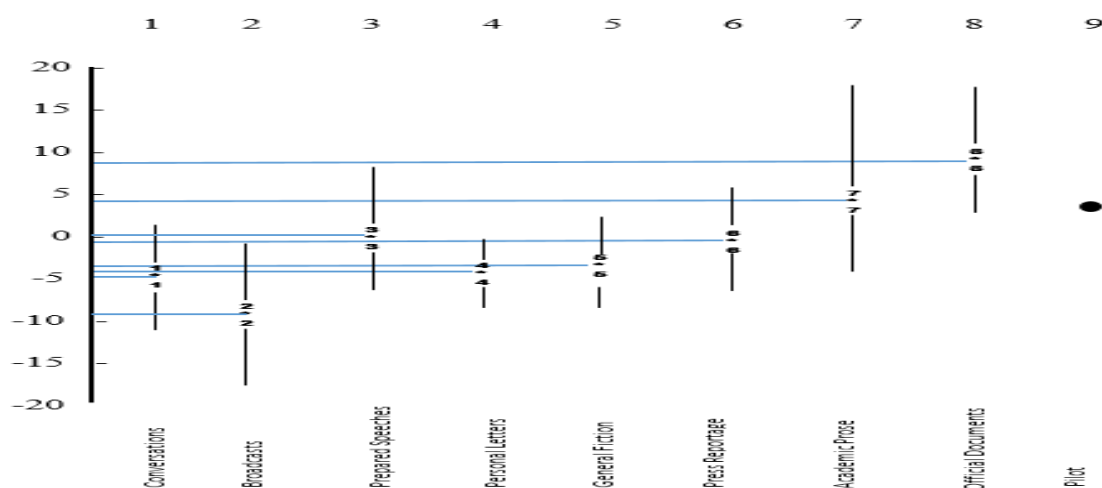


Figure 2: Cross Mean Value

5th genre of general fiction is also not much prominent in the pilot test of the data with reference to the 3rd linguistic dimension of explicit vs situational dependent reference. For the reference, press reports are founded in the data but the overall mean stay in the negative y-axis in the 6th stage of genre but it is also near to absolute zero value. 7th genre of adamic prose is the most prominent and dominant mean value that effected the results and consider the twitters data as a formal source of information with a slight oscillation of informal vocabulary identification. Twitter is also used by officials of political parties and mostly the governing bodies post their public documents on twitter along with the confirmation of authenticity. Due to this reason, the mean value of official documentation is near about 10 on +y-axis. The mean configuration of stander aviation is 2.46 in the pilot test file of twitter data. Sketching the oscillatory statistical three ad, a lower to higher pointer connector visualization is appeared which shows the true reliability of selected linguistic dimension for the data analysis. With the indication of differential oscillation on the output results, the procedure of statistics are more illuminated and make the linguistic dimensional results prominent. This rectification is necessary to display on the data with using the line tool for the understanding of genre analysis on graded scale.

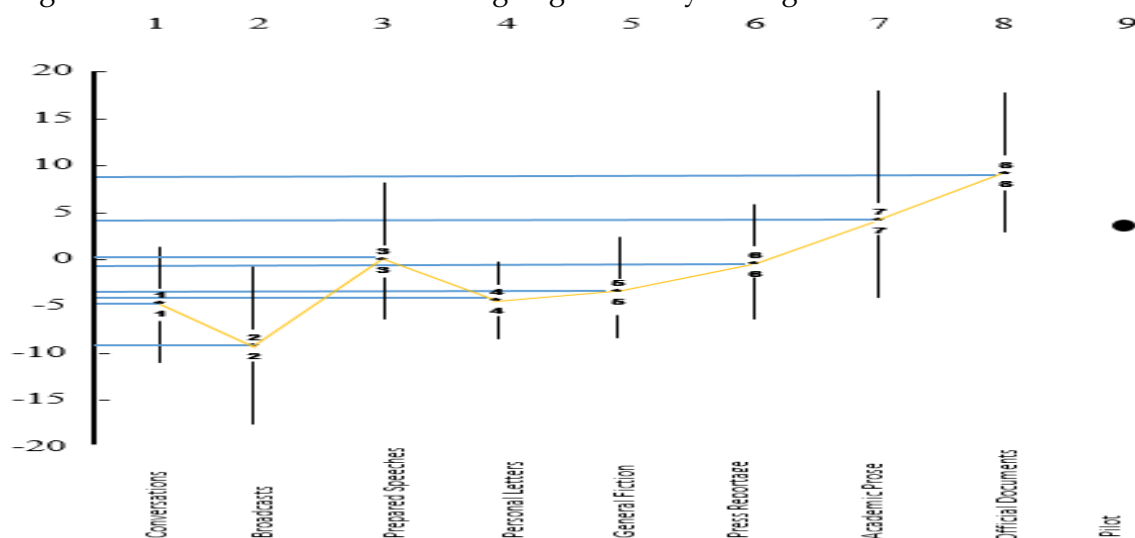


Figure 3: Differential Oscillation Among Genres

Figure 4 is a screen clipping of reference corpus dimensional result which is figured out by Biber. The mean value of reference 3rd dimension is 4.9 which is .3 times lesser in extracted results with 4.6 as mean value of 3rd dimension. As Halliday elaborated the level of formality and informality in speech so the statistics results shows the true variation between explicit and implicit vocabulary used by the social media users of twitter in political argues

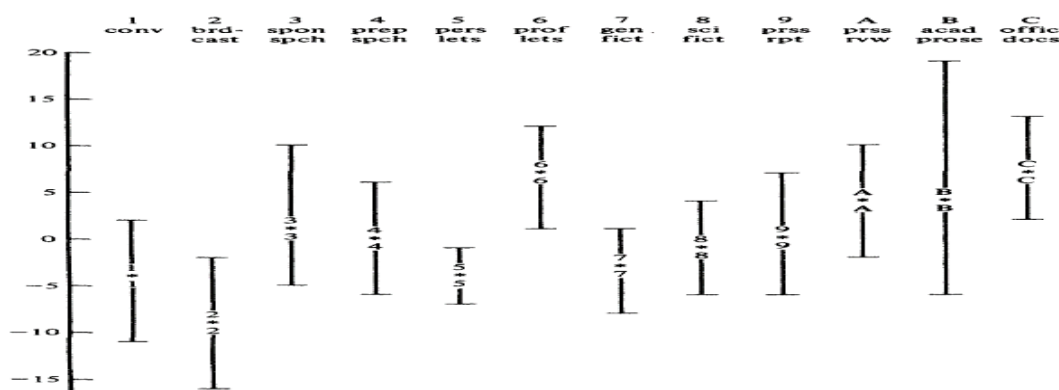


Figure 4: Reference Corpus 3rd Dimensional Analysis

**Table 1**  
**Reference Pilot**

| Statistics   | Dimension                           | Mean | Minimum Value | Maximum Value | Range | Standard Deviation |
|--------------|-------------------------------------|------|---------------|---------------|-------|--------------------|
| Reference    | Dimension 3-                        | 4.9  | -0.4          | 10.3          | 10.6  | 3.6                |
| Pilot        | Explicit vs                         | 4.6  | -0.8          | 9.6           | 9.87  | 2.46               |
| Differential | Situation<br>Dependent<br>Reference | 0.3  | -0.4          | 0.7           | 0.73  | 1.14               |

### Dimension 3- Explicit vs Situation Dependent Reference

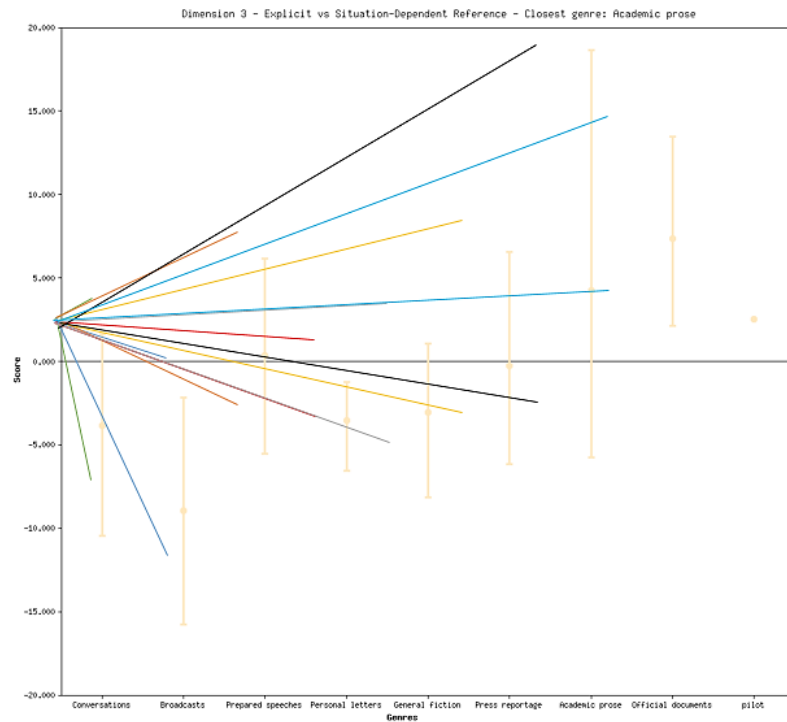


Figure 5: Mean Comparison of Genres with Absolute Zero

Total number of tokens that are tokenized with MAT tagger in the data are 990289 from which academic word list is 6.92 percent of the total text. There are 570-word families in AWL generated by Averil Coxhead, serve the purpose of tagging the academic vocabulary words in the processed corpus. Type token ratio “TTR” indicate the static of 268 differentials in the political text of twitter to show the lexical diversity and writing style. LRB and RRB remain constant in the value consideration of 0.7 concentration in the text. Amplifiers are founded only 0.07 percent in the English language text because the mixing of Urdu language may legitimize the intensity and count of AMP.

**Table 2**  
**Statistics Pilot**

|        |        |      |       |        |      |
|--------|--------|------|-------|--------|------|
| Tokens | 990289 | JJ   | 3.02  | VBN    | 0.62 |
| AWL    | 6.92   | LS   | 0.01  | VPRT   | 3.97 |
| TTR    | 268    | NEMD | 0.12  | WDT    | 0.02 |
| #      | 0.01   | NN   | 31.13 | WP     | 0.19 |
| \$     | 0.01   | NNP  | 0     | WPS    | 0.01 |
| "      | 0.5    | NOMZ | 1.44  | XX0    | 0.52 |
| -LRB-  | 0.07   | OSUB | 0.02  | [BEMA] | 1.2  |
| -RRB-  | 0.07   | PDT  | 0     | [BYPA] | 0.01 |
| .      | 3.15   | PHC  | 0.42  | [CONT] | 0.55 |
| :      | 3.99   | PIN  | 4.73  | [PASS] | 0.12 |

|                 |              |          |       |          |       |
|-----------------|--------------|----------|-------|----------|-------|
| AMP             | 0            | PIT      | 0.35  | [PASTP]  | 0     |
| A.S.            | 0            | PLACE    | 0.11  | [PEAS]   | 0.25  |
| AMP             | 0.07         | POMD     | 0.18  | [PIRE]   | 0     |
| ANDC            | 0.04         | POS      | 0.3   | [PRES P] | 0.05  |
| CAUS            | 0.12         | PRED     | 0.31  | [PRIV]   | 0.98  |
| CC              | 1.04         | PRMD     | 0.32  | [PROD]   | 0.04  |
| CD              | 2.38         | PRP      | 0.06  | [PUBV]   | 0.26  |
| COMMA           | 1.55         | QUAN     | 0.5   | [SERE]   | 0     |
| CONC            | 0            | QUPR     | 0.07  | [SMP]    | 0.02  |
| COND            | 0.21         | RB       | 1.38  | [SPA U]  | 0.12  |
| CONJ            | 0.05         | RP       | 0.22  | [SPIN]   | 0     |
| DEMO            | 0.27         | SPP2     | 0.74  | [STPR]   | 0.01  |
| DEMP            | 0.12         | SYM      | 0.12  | [SUAV]   | 0.44  |
| DPAR            | 0.02         | SYNE     | 0.43  | [THATD]  | 0.18  |
| DT              | 2.5          | THAC     | 0.01  | [WHCL]   | 0.02  |
| DWNT            | 0.05         | THVC     | 0.08  | [WHOBJ]  | 0.01  |
| EMPH            | 0.13         | TIME     | 0.39  | [WHQU]   | 0.02  |
| EX              | 0.11         | TO       | 0.91  | [WHSUB]  | 0.09  |
| FPP1            | 1.67         | TOBJ     | 0.1   | [WZPAST] | 0.02  |
| FW              | 32.04        | TPP3     | 1.39  | [WZPRES] | 0.18  |
| GER             | 0.16         | TSUB     | 0.05  | _NN      | 0.09  |
| HDG             | 0            | UH       | 0.01  | _VPRT    | 0     |
| IN              | 0.47         | VB       | 2.47  | ``       | 0.3   |
| INPR            | 0.02         | VBD      | 1.25  | ga.      | 0     |
| Tokens          | 990289       | VBG      | 0.87  | http     | 0     |
| <b>Filename</b> | <b>pilot</b> | PLACE    | -0.59 | [PRIV]   | -0.79 |
| AMP             | -0.77        | POMD     | -1.14 | [PROD]   | -0.74 |
| ANDC            | -0.85        | PRED     | -0.62 | [PUBV]   | -0.94 |
| AWL             | 6.05         | PRMD     | -0.57 | [SERE]   | -0.25 |
| CAUS            | 0.06         | RB       | -2.94 | [SMP]    | -0.6  |
| CONC            | -0.63        | SPP2     | -0.18 | [SPA U]  | -1.72 |
| COND            | -0.18        | SYNE     | 1.63  | [SPIN]   | 0     |
| CONJ            | -0.44        | THAC     | -0.33 | [STPR]   | -0.7  |
| DEMO            | -1.71        | THVC     | -0.86 | [SUAV]   | 0.48  |
| DEMP            | -0.71        | TIME     | -0.37 | [THATD]  | -0.32 |
| DPAR            | -0.43        | TO       | -1.04 | [WHCL]   | -0.4  |
| DWNT            | -0.94        | TOBJ     | 0.18  | [WHOBJ]  | -0.76 |
| EMPH            | -1.19        | TPP3     | -0.71 | [WHQU]   | 0     |
| EX              | -0.61        | TSUB     | 0.13  | [WHSUB]  | -0.6  |
| FPP1            | -0.4         | TTR      | 3.06  | [WZPAST] | -0.74 |
| GER             | -1.42        | VBD      | -0.91 | [WZPRES] | 0.11  |
| HDG             | -0.46        | VPRT     | -1.11 |          |       |
| INPR            | -0.6         | XX0      | -0.54 |          |       |
| JJ              | -1.62        | [BEMA]   | -1.72 |          |       |
| NEMD            | -0.43        | [BYPA]   | -0.54 |          |       |
| NN              | 3.67         | [CONT]   | -0.43 |          |       |
| NOMZ            | -0.38        | [PASS]   | -1.27 |          |       |
| OSUB            | -0.73        | [PASTP]  | -0.25 |          |       |
| PHC             | 0.3          | [PEAS]   | -1.17 |          |       |
| PIN             | -2.49        | [PIRE]   | -0.64 |          |       |
| PIT             | -0.96        | [PRES P] | -0.29 |          |       |

## Conclusion

Concluding the whole descriptive analysis on the formal and informal register variation it is evident that., the level of linguistic dimension and genre is always based on



the features of the text. Twitter is a social media platform where mostly the political constrains are discussed and using the English language in Pakistan is a considered as a power social factor. Due to this reason, the data statistics stay more closer to the formal register and highlighted as political adamic prose. The total number of tokens applied on the text are statistically included in the result bar of Table 1. Every expression of English Language is encountered in the Static Pilot Test but due to the Urdu text inclusion, results may vary accordingly. Hence the dissection of the whole data is drawn with the point-to-point analysis of linguistic hierarchy with the referential terms indicated in the reference corpus.

### **Recommendations**

Considering the delimitations of the study, all dimensions are not incorporated to make focus on only 3<sup>rd</sup> dimension analysis of MAT tagger. Future researcher with the same extension may corporate the other dimensions and theoretical approach to explore

## References

- Abbas, H. (2018). *Linguistic Variation Across Pakistani English Written Registers: A Multidimensional Analysis*. Department of Applied Linguistics, Government College University, Faisalabad, Pakistan.
- Ali, S. H. (2019). A corpus-based analysis of the use of formal and informal vocabulary in Pakistani English on Twitter. *Journal of English as an International Language*, 14(2), 16–34.
- Aziz, S., & Ali, S. (2018). Formality and informality in Pakistani English tweets: A corpus-based analysis. *GEMA Online Journal of Language Studies*, 18(4), 149–165.
- Buckingham, D. (2008). *Youth, identity, and digital media*. MIT Press.
- Halliday, M. A. K. (1960). *Language as social semiotic: The social interpretation of language and meaning*. Edward Arnold.
- Khan, N. A., & Ahmed, Z. (2020). Investigating the use of formal and informal vocabulary in Pakistani English on Twitter. *Language in India*, 20(10), 148–162.
- Lukin, A., Moore, A., Herke, M., Wegener, R., & Wu, C. (2011). Halliday's model of register revisited and explored. *Linguistics and the Human Sciences*, 4(2), 187–213.
- Mahmood, R. (2009). *A Corpus-Based Analysis of Pakistani English*. Department of English Language & Literature, Bahauddin Zakariya University, Multan, Pakistan.
- Maybaum, R. (2013). *Language change as a social process: Diffusion patterns of lexical innovations in Twitter*. *Proceedings of the Berkeley Linguistics Society* (Vol. 39, pp. 152-166). Linguistic Society of America.
- Raza, S., & Tariq, S. (2019). A corpus-based study of formal and informal vocabulary in Pakistani English tweets. *International Journal of Linguistics, Literature and Culture*, 5(3), 56–69.
- Shaheen, R., & Farooq, U. (2018). Formal and informal vocabulary use in Pakistani English tweets: A corpus-based study. *Journal of Language and Literature*, 9(1), 1–12.