Performing Natural Language Processing on Roman Urdu Datasets

Zareen Sharf and Dr Saif Ur Rahman

PhD Scholar SZABIST

Summary
This work is a predecessor of a larger task which requires discourse based sentiment analysis on Roman Urdu Datasets. In order to perform this task, we first needed to collect a large data corpus in Roman Urdu from social Media websites. Next we cleaned the raw data, lexically normalized it for standard representation of words, performed POS tagging for the words to be tokenized meaningfully and finally identified the presence or absence of a discourse element. After achieving these task, we are now ready to perform Neural Network based sentiment Analysis on Roman Urdu dataset taking discourse into consideration as our future work.

Key Words:
Natural Language Processing, POS Tagging, Discourse units, Roman Urdu Data

1. Introduction

An essential phenomenon in natural language processing is the use of discourse relations to establish a coherent relation, linking phrases and clauses in a text. The presence of linguistic constructs like connectives, modals, conditionals and negation can alter sentiment at the sentence level as well as the clausal or phrasal level. Consider the example, "@user share 'em! I'm quite excited about Tintin, despite not really liking original comics. Probably because Joe Cornish had a hand in." The overall sentiment of this example is positive, although there is equal number of positive and negative words. This is due to the connective despite which gives more weight to the previous discourse segment. Any bag-of-words model would be unable to classify this sentence without considering the discourse marker.

Consider another example, "Z10 Kaafi Intresting Set laga Lekin me bettry timing se thora dar gya hon overall set acha he Lekin 20hzaar Is set pe kharch karna Kia sahe he ya koi 20hzaar tak ka set jo apki nazar me ho jis ki ram nd storage healthy ho Ar bettry timng bi achi dy agar ap bta dein to acha hoga" The overall sentiment neutral due to the connective but, which gives more weight to the following segment of the comment. Thus it is of utmost importance to capture all these phenomena in a computational model.

Our focus is mainly on developing a discourse parser for Roman Urdu text so we can perform discourse based opinion mining on text. We intend to exploit the various features discussed in the Twitter specific works to develop a model, in which the discourse features are incorporated to give better sentiment classification accuracy.

2. What is discourse?

Discourse in terms of natural language is an essential paradox in natural language processing. Discourse helps in authorizing coherent relation, phrase and clauses linking within a text segment (Mukherjee, 2012). Discourse can also be defined as a logical structured group of textual units or segments. A discourse can be in any form be it a sentence, a dialogue, a written text and etc. Discourse coherent structures specify relations between two sentences or clauses. Discourse can cause two sentences to be coherent or related. (Computational Discourse, n.d.) A connective supports the discourse relation when it comes in between two segments. A connective gives weight to either side of the two connected sentences giving a meaning to the sentence be it positive or negative.

Some theories exist to analyze discourse and the segments involved in classifying a discourse. Some of these theories are Rhetorical Structure Theory (RTS) which was proposed by Mann et al. (1988). This theory tries to idealize the two segments called nucleus and satellite in sentence. Much work has also been done in establishing elementary discourse units at clausal stage and by generating trees for sentence level. A discourse parser or a dependency parker is used in most of these discourse based works. (Mukherjee, 2012)

3. What is discourse analysis?

Discourse analysis can be understood as the study of language in texts and conversation. Critical Discourse Analysis is the branch of linguistics that deals with understanding why and how some writings have more of an impact on readers and hearers than some others. Even by the analysis of simple grammar, the aim of critical discourse analysis is to find out the ideologies that are

Manuscript received January 5, 2018
Manuscript revised January 20, 2018

hidden and have the abilities to influence the view of the world of the reader and the hearer. Discourse analysts have analyzed a wide range of writings and texts and even spoken manifestos and rules, in order to show how writers and speaker can become influential through their words and become ideologically significant.

In this era, micro-blogs have created an impact on the society. Hence it was important to discuss the impact that the micro-blogs have on human perception. A micro-blog is one that allows small content such as small sentences and video links to be shared among people. Certain theories have been proposed in order to analyze micro blogs specifically. But there are certain problems that arise in the implementation of these theories. For example, when talking about twitter, the tweets that people make are not restricted by the content that the user uses or the form it takes. Normally people who tweet, are not confined to using formal language and there isn't a lack of spelling mistakes as well or even the use of discontinuities and even grammatical errors. Due to this, natural processing tools for language such as taggers and parsers fail because of their inability to handle unstructured data. (Dey, Lipika and Haque, Sk., 2009)

Normal language processing tools make use of modals, connectives and conditionals to analyze the text but when talking about micro-blogs, these formalities are mostly ignored and are replaced by domain-specific specialized characteristics and features such as the use of hashtags and emoticons. (Alec, G.; Lei, H.; and Richa, B., 2009)

For the identification of elementary discourse units in order to generate trees at the level of sentence, (Marcu, Daniel, 2000) proposed probabilistic models. This model used syntactic information and lexical information from discourse-annotated corpus. The effect of negatives, modals and connectives changing the prior polarity of the said words in order to make out new meanings was investigated in Contextual Valence Shifters by Polanyi in 2004. These talk about a weighting scheme that is simple and also talk about pre-suppositional items and irony.

When talking specifically about analyzing the discourse in micro-blogs, the following studies come up. The most commonly used feature on twitter is the hashtag. (Alecel al., 2009), used the hashtags in the tweets that people made, to make a training data set in order to perform a multi-class classifier with the use of clusters that were topic dependent. This was the approach proposed for sentiment classification using a distant supervision-based approach.

(Joshi et al, 2011), proposed a rule that classified tweets either as negative or positive depending upon the specific opinion words present inside it. In order to classify, it made use of twitter specific features such as hashtags, emoticons and sentiment lexicons. The use of emoticons and hashtags is so common on social media that it is necessary to incorporate them in any discourse analysis rule. In order to distinguish sarcasm, wit and negative and positive tweets, Gonzalez in 2011 also relied on the information provided in the hashtags. According to him, the hashtags are the best indicators of whether the said discourse is sarcastic or not.

A recent study that is based upon the works of (Wolf, Florian and Gibson, Edward, 2005)and (Taboada et al. , 2008) (Polanyi et al. , 2004). The study furthers in sentiment analysis of these micro blogs. In this particular study, the features that are discussed in works specific to twitter are exploited in order to develop a bag-or-words type model. In this model, in order to give a better accuracy of sentiment classification, features relating to discourse are incorporated. Three sets of data are taken and evaluated by making use of classification that is lexicon-based and also making use of supervised classifier. In this study, labelled tweet sets are used manually with more than 8500 tweets and another set that is automatically annotated containing almost 15200 tweets. In this study, further datasets were made use of from the travel review domain of 2011 by Balamurali et al. This was incorporated in order to show the method employed was beneficial to reviews that were structured as well. (Mukherjee, 2012)

4. Types of Discourse

Ordinary language discourse mostly appears as the expression of emotions, feelings or attitudes. It can be therefore classified as follows.

4.1 Coherently Structured Discourse

A group of sentences having some relationship with each other is called coherently structured discourse. Their relation is explained by a coherent relation and how they interact with each other. The coherent relation between sentences in a discourse structure differs with respect to two separate approaches. In one approach aims to equate intentional level structure of discourse. In this approach the coherence relation imitate how one segments role played with respect to interlocutor's purpose communicates to the another segment's role. (Grosz, Barbara J. and Candace L. Sidner, 1986). The other approach aims to idealize a discourse's informational structure. In this approach the coherence relations imitate how meaning transmitted by one discourse segment divulge with the meaning transmitted by other discourse segment. (Hobbs, Jerry R., 1985) (Marcu, Daniel, 2000).

Conjunctions used to illustrate Coherence Relations in Discourse (Wolf, Florian and Gibson, Edward, 2005):

The coherent interaction between to discourse segments can be of various types such as:

- Cause-effect because; and so
- Violated Expectations although; but; while
- Condition if…(then); as long as; while
- Similarity and; (and) similarly
- Contrast by contrast; but
- Temporal Sequence (and) then; first, second, … before; after; while
- Attribution according to …; …said; claim that …; maintain that …; stated
- Example for example; for instance
- Elaboration also; furthermore; in addition; note (furthermore) that; (for, in, with) which; who; (for, in, on, against, with) whom
- Generalization in general

4.2 Explicit Discourse

A discourse is termed explicit when the text does not contain any explicit cues. This type of discourse is signaled by connectives likes however, since, because and etc. (Syeed Ibn Faiz and Robert E. Mercer) Explicit discourse relations are very easy to identify. Comparison, contingency, temporal and expansion, these general senses can easily be authorized in explicit discourse relations with about 93% of accuracy. This accuracy is based entirely on the usage of discourse connective to signal relation. (Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi., 2008). To identify explicit connectors Penn Discourse Treebank is used usually for experimenting. It is a largely illustrated bulk collection of discourse relations. (Prasad et., al, 2008). Other samples usually used for explicit discourse relation analysis is English Giga-word Corpus which contains above four million news articles. (Graff, 2003)

4.3 Implicit Discourse

When there exist a discourse connective between two text segments, it usually easy to identify the actual relation between these segments, as they contains connectives that are unclear (Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber., 2005). On the contrary it gets difficult to recognize relations where no explicit textual cues are found. (Marcu & Echihabi, 2002)Worked for the first time to detect implicit discourse relations. They classified implicit discourse relation by showing that words summarized from two text segments to detect implicit discourse relation between text segments. In implicit discourse the discourse relation does not have any connectives and two text segments are mostly adjacent to each other. (Ziheng Lin, Min-Yen Kan & Hwee Tou Ng)

5. Literature Review

(Joty, Shafiq, Giuseppe Carenini, and Raymond T. Ng, 2012.) Proposed a framework for discourse parsing at sentence level which was based on complete probabilistic discriminative structure. This framework was composed of a discourse segmenter which was based upon a binary classifier along with a discourse parser which parsed probabilities that were extracted from a Dynamic Conditional Random Field using an optimal CKY like algorithm for parsing. The data corpus was divided into a training set containing 347 documents, a test set containing 38 documents and 53 human annotated documents. The results were compared with the results of HILDA, SPADE and the results reported in Fisher and Roark (2007). Even by using fewer features the results of the given approach outperformed other parsers by a wide margin. The weakest performance was given by HILDA. The absolute F score was computed to be 4.9 percent. The discourse parser could be improved by better representation of semantic knowledge. A more robust methodology is required for the imbalanced distribution of relations. This framework can be extended using graph structures for discourse analysis. This parser can be generalized to multi sentential text to verify the limit to which segmentation can be jointly performed with parsing. (Mitocariu, Elena, Daniel Alexandru Anechitei, and Dan Cristea, 2013) Presented three new scores for the comparison of discourse trees. These scores took additional constraints into consideration. Two discourse theories were used to build the discourse structure: Rhetorical Structure Theory (RST), Veins Theory Most of the existing discourse parser use Precision, Recall and F score for comparing the discourse trees. These scores can be only used if the topological structure is identical. Therefore, in order to compare discourse trees three scores were proposed: The Overlapping Score (OS) which took into consideration the coverage of nodes only. The Nuclearity Score (NS) which considered the nuclearity of relations and the Vein Scores that computed the F score, Recall and Precision of the elementary discourse units. The first two measures were beneficial for the comparison of discourse tree structures while the third one was significant for summarization applications. These scores compare different parsers efficiently as the non-relevant idiosyncrasies are not noticed. They are most useful for summarization where main ideas are recognized by nuclearity of discourse units.

(Song, Wei, et al., 2015) discuss identification of discourse elements using sentences of persuasive essays written by Chinese students of high school. The proposed method used cohesion to improve the identification of discourse elements. Cohesion was defined as a group of resources that linked the text together. Like use of conjunction, substitution, reference, ellipsis and lexical

cohesion. In this study the local and global cohesive relation were examined. Sentence chains were created based on cohesive resources and were then examined whether they represented local or global cohesion. For the functional segmentation of discourse, machine learning models were used. Two representative models (SVM and linear-chain CRF) were used for evaluation. The results supported the hypothesis that by adding the cohesion features a significant improvement could be attained. The F1 scores were all found to be significant with a p value less than 0.01 using the pair wise student t-test. All discourse elements of the 3 corpora showed improvement with the addition of cohesion features. When the confusion matrix was analyzed it was found that this improvement was mainly the result of the distinction between the thesis and main idea sentences. The chain related features alone did not show considerable discriminative ability. But when they were combined with the cohesion features the F1 score rose by 0.9 percent.

(Lüngen, Harald, et al., 2006) describes discourse segmentation as the decomposition of text into small sections. This paper explained the process of discourse segmentation by using Segmentation principles, morphology, punctuation and syntax for German language. An automatic Rhetorical Structure Theory (RST) based discourse segmenter was developed. It evaluated how the discourse units were defined and recognized automatically by using principles of segmentation for English elementary discourse units. As training data was not available in huge quantity for German Language the discourse segmenter used a knowledge based procedure. A syntactic parser had been integrated that worked online in the process of segmentation. A corpus based on forty-seven scientific German articles were used to develop the discourse parser and segmenter. In addition, a newspaper as well as a web published article were also segmented to further evaluate the performance of the segmenter on other types of texts. The EDS segmenter performed significantly better in comparison to the baseline version for all of the six texts. However, the EDS along with the SDS segmenter performed slightly worse than the segmenter for English. The recall figures were also lower than that of the reported Statistical approach for discourse segmentation with an overall recall of 79 percent. Some errors were produced while the implementation of the EDS segmenter. These errors were due to the recognition failure of the attributional construction and the defective analysis in the complex and long sentences done by the syntactic parser. Both omission as well as too much use of commas was another reason that produced errors. However, these errors could be resolved with the help of syntactic analysis which could improve the performance in future.

(Webber, Bonnie, and Aravind Joshi., 2012) reported the intrinsic features of discourse and its properties. The study highlighted that 4 types of discourse structure had made progress. These were:

1. Topic structure that deals with breaking a discourse into sequence of topics
2. Functional structure identifies sections having different functions within a discourse
3. Event structure is a recent phenomenon that deals with identification of events.
4. Structure of coherence relations deals with discourse relations such as contrast and succession and condition and motivation.

The study further discussed the resources that were employed to recognize and label these structure. The first issue dealt with the evidence for a specific discourse structure. The variability in the annotation of discourse structure was another issue. Furthermore, when machine learning methods were adopted the speaker intentions and centrality of pragmatics were abandoned. There were no reliable proxies left. These data intensive methods also eliminated the issues and inferences related to implicit information. Greater openness in conveying information and better modelling techniques could resolve the issues. If the system performance in recognizing the roles played by utterances was improved for one genre it could be generalized and transported between genres. Further research could be needed to exploit discourse for understanding the inter dependence of different features of discourse structure, widen the gained knowledge, information extraction and to continue the discourse research in multiple languages.

(Forbes-Riley, Kate, Fan Zhang, and Diane Litman, 2016) Examined the automatic and manual annotation of Penn Discourse Treebank (PDTB) discourse relations using English essays written by school students. The methodological complexities that were required for this automatic annotation have also been discussed and its performance has been compared to prior work.

This study used the Penn Discourse Treebank (PDTB) agenda to add the annotation of discourse relation to numerous corpora. It focused on the lexical basis of discourse relations. The corpus used in this study to examine the annotation differed significantly from the ones used before. The essays were learning based as through the process of essay writing, the writers learned argumentative writing. These essays also contained various spelling and grammar mistakes as well as cohesive issues. The study postulated that these differences would explain the unclear features of the PDTB approach and would possess a challenge for an automatic discourse parser. Descriptive statistics of PDTB and BioDRP corpus were compared. The text of BioDRP were not learning based and they did not possess an argumentative structure. To predict the human annotated relation the Lin et al.

parser which was PDTB trained was used. It was claimed to be the first end to end free text discourse parser based on PDTB framework. First it identified the discourse connectives and then assigned a sense after identifying their two arguments. This parser has been used in 2 different ways. First the parser is used at Level 2 so that the essays can be parsed according to Level 2 senses then the Level 2 senses are converted Level 1 abstractions. Secondly, the parser was retrained at Level 1 senses which predicted the Level 1 senses directly.

By comparing the performances of discourse parsers it was concluded that the negative effect of noise is significantly eliminated when the minimal argument constrain is relaxed and only level-1 senses is predicted. The Lin et al. parser resulted in a F1 score of 31 percent which is similar to other parsers. The performance was highest during the identification of argument and connectivity and fall rapidly during the identification of sense and relation type. The parsers ability to differentiate Implicit/Expansion, EntRel and AltLex can be improved by training on essay data. The results were in accordance to prior work and are in favor of modifying the manual annotations according to the target data and training the domain-specific parsers for prediction.

6. Parsing and Tagging of Roman Urdu Datasets

According to (ABBAS, 2015) parsing is the division of sentences into grammatical parts including identification of parts of speech and relationship of each word segment to each other. Many treebanks and parsers have been developed but they are not capable of parsing Treebanks for morphologically rich languages like Urdu, Italian, French etc. Urdu is a less modernized language for which a development of a modern Treebank and a parser would be beneficial in modern automatic language processing. The solution provided in this paper is in terms of the development of a parser and a rich Treebank for the Urdu language. For development of Urdu parser, 1400 annotated sentences of URDU.KON-TB were divided in 80% training data and 20% test data. And following steps were followed:

1. Context free grammar was concluded from training data and given to parser for development.

2. 10 % held out data and 10% test data is the division of test data.

3. The sentences in test data consist of an average length of 13.73 words in a sentence.

4. The held back data is used in parser development.

The Urdu parser developed was an extended version of the Earley parsing algorithm. This research produced semi-semantic syntactic tagset, annotation guidelines, grammar sufficient encoding for MRL Urdu language, and semi semantic part of speech tagging. However, the problem still lied in the annotation of URDU.KON-TB Treebank and annotation guidelines. Also the difference in reported values of annotation evaluation and parsing evaluation can be improved further.

(Tafseer A., Saba U., Sarmad, H, Asad M., Rahila P. Farah Adeeba, Annette Hautli, Miriam Butt.) focused on developing a tagger for the Urdu Language. The tagger which was analyzed in this paper faced problems in comparing same pair of tags, as in Urdu there are two tags for Nouns i.e., Noun and Proper Noun, although in Urdu Language there is no clear distinguishing between these two types of nouns. Nouns are also confused with adjectives when they are placed side by side. To overcome the problems mentioned in Problem statement a new CLE Urdu POS Tagset is introduced. This tagset used tags that include and incorporate information from special morpho-syntactic categories found in Urdu Language. The naming schema and basic divisions are done according to Penn Tree bank and Common Tagset for Indian Languages. The new tagset is formed by doing three main functions: Comparison with other available tagset, Linguistic issues and syntactic distribution and Pre-tagging of 100k words of CLE Urdu Digest balanced corpus. The tagset was proposed mainly by giving new tags that are analyzed by a morpho syntactic pattern. The CLE Urdu Tagset proposed in this paper concludes 12 categories and results in 32 tags. The tagset tagged 100k words in which 80% was training corpus and 20% was testing corpus and the files to be analyzed were selected randomly. The Tree Tagger was used to do automated tagging and a Decision tree and smoothing technique of Class Equivalence was used as machine learning technique. This resulted in successful tagging of 100k words with accuracy of 96.8%.

(ABBAS, 2015) proposed an automated parser for morphologically rich Urdu Grammar. The URDU.KON-TB was used for annotated source and the automated parser was developed my modifying the Earley algorithm. The automated parser worked with a sentence having length of five tokens for which the parser generated 0 to 5 charts. All unused NL type segments were removed. The charts were then regenerated through the parser.

The Urdu parser was compared with other language parsers such as Hindi parser in which other parser performs with 22% recall. Urdu Parser was also compared with MPSRP in which 78% accuracy was achieved by Urdu Parser. The parser proved to be efficient in parsing due to usage of URDU.KON-TB as a grammar parser. The parser was capable of correcting itself. It could edit

and correct failed parsed output as well. The Treebank contained around 1400 annotated sentences which could be increased in future. The Urdu Parser does not cover unknown sentences in partial parsed trees which is a limitation in adding more annotated sentences in the bank. This could be done once partial parsed trees got corrected and imported to URDU.KON-TB Treebank.

7. Methodology

According to (Tafseer A., Saba U., Sarmad, H, Asad M., Rahila P. Farah Adeeba, Annette Hautli, Miriam Butt.) POS tagging is a process of analyzing a sentence and then assigning parts of Speech to each word of the sentence. POS tagsets are the tags used to incorporate defined tags to words of a language. A tagset must be smart enough that it can encode grammatical differences of interest so that can be used in machine learning, natural language processing etc. For Urdu many POS tagsets have been developed that use CLE Urdu Digest corpus as it is the only largely available corpus for Urdu language but no such resource was available for Roman Urdu Corpus. Therefore, in order to perform NLP on Roman Urdu our first challenge was collection of a significantly large data corpus.

Our research methodology was based on quantitative research methods as our focus was to generate results from the collected data corpus. The statistical evaluation of results has been done on F1 Score method. We started off with scraping data using Python coded scripts from websites like Bio Social Workers, Bio Graphies, Blog Khuwaar, Reddit, City News Tweets, Express Urdu Tweets, Nida Imranist, Urdu SMS, Shashca and Pakish News. Next step involved preprocessing and cleaning of raw data. This step was followed by lexical standardization of retrieved data which we performed using our own algorithm that used the same principle on which Soundex is based. Lexical normalization is the process of unifying the representation of a word that might be spelled differently by different entitites into a single form. The third step involved creating a word list from all the word that were present in our corpus. These we stored in order of frequency of their appearance in the source dataset. Once the wordlist had been generated we then manually annotated the POS tag for each entry taking help from Google and IJunoon Transliteration services.

The POS module works with the pre-tagged wordlist compiled from the data corpus. The module however stores any word that is not included in the list as a separate list which is then manually annotated by human expert and added to the wordlist. In this way new words keep getting appended as the data corpus increases. The discourse functionality also works in sync with this module where words like “agar, magar, laiken, halankay, baharhaal, kyunkay, chunkay” etc are pretagged to represent discourse elements so this is also identified by the POS module. The final step involved the identification of presence of a discourse element in the input sentence.

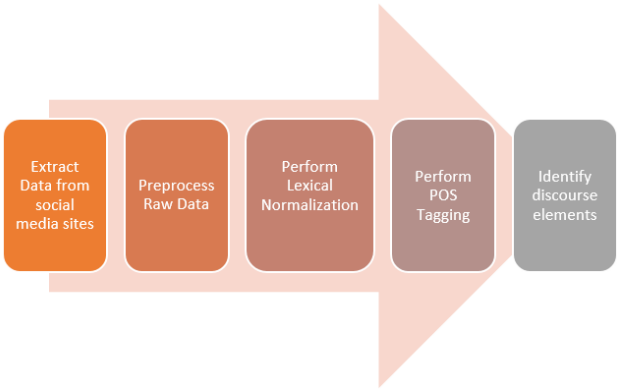

Fig. 1 System Data Flow Diagram

8. Data Analysis

We used a dataset comprising of more than 15000 statements in Roman Urdu. F1 Score calculation was used for evaluating the accuracy of the results. In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

We have analyzed data from different websites namely, Twitter, Reddit, Urdu Poetry and Social workers Biographies. The tabulated results achieved are shown in Table 1.

Table 1: Dataset for Discourse based POS Tagging

S#	Source	Number of Sentences	Number of Sentences with discourse	Correct	Wrong	Success Rate	False negative	False positive
1	Bio Social Workers	1500	250	195	55	78%	20	35
2	Bio Graphies	15000	1400	1260	140	90%	35	105
3	Blog Khuwaar	350	100	87	13	87%	5	8
4	Reddit	150	25	23	2	92%	0	2
5	City News Tweets	1300	140	111	29	79%	12	17
6	Express Urdu Tweets	2700	200	158	42	79%	3	39
7	Nida Imranist	250	30	19	11	63%	2	9
8	Urdu SMS	625	51	45	6	88%	0	6
9	Shashca	62	7	7	0	100%	0	0
10	Pakish News	125	14	12	2	85%	0	2

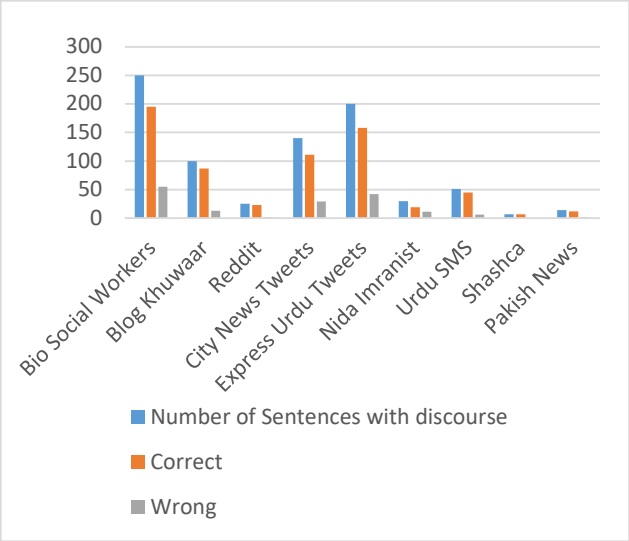

Fig. 2 Comparative Analysis of Discourse based POS Tagging

As per definition (Wikipedia)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{F1 Score} = 2 \times \frac{\text{precision x recall}}{\text{precision + recall}}$$

Table 2: F1 Score Parameters Matrix

		Prediction Positive	Prediction Negative
True Condition	Condition Positive	True Positive	False Negative
	Condition Negative	False Positive	True Negative

Table 3: F1 Score Values from the given Datasets

S#	Source	Precision	Recall	F1 Score
1	Bio Social Workers	0.84	0.906	0.871
2	Bio Graphies	0.92	0.97	0.944
3	Blog Khuwaar	0.91	0.94	0.924
4	Reddit	0.92	1	0.958
5	City News Tweets	0.86	0.902	0.88
6	Express Urdu Tweets	0.802	0.98	0.882
7	Nida Imranist	0.67	0.904	0.769
8	Urdu SMS	0.88	1	0.936
9	Shashca	1	1	1
10	Pakish News	0.85	1	0.918

9. Results

The results reveal a high value of F1 score for the dataset comprising of data from ten different sources. The reading from one of the datasets is low approximately 68% while other datasets show an accuracy of around 80% on an average. The possible reason for a low value could be the incorrect standardization of words that are consequently missed by the tagger. These types of shortcoming can be rectified by tuning the transformation rules in the lexical

normalization module and also by producing a bigger set of words for better POS tagging capabilities.

10. Conclusion

This work is a predecessor of a larger task which requires discourse based sentiment analysis on Roman Urdu Datasets. In order to perform this task, we first needed to collect a large data corpus in Roman Urdu from social Media websites. Next we cleaned the raw data, lexically normalized it for standard representation of words, performed POS tagging for the words to be tokenized meaningfully and finally identify the presence or absence of a discourse element. After achieving these task, we are now ready to perform Neural Network based sentiment

References

[1] ABBAS, Q. (2015). Morphologically rich Urdu grammar parsing using Earley algorithm. Natural Language Engineering, 1-36.

[2] Abbas, Q. (n.d.). Building Computational Resources: The URDU.KON-TB Treebank and the Urdu Parser. Fachbereich Sprachwissenschaft, Fachbereich Sprachwissenschaft.

[3] Alec, G.; Lei, H.; and Richa, B. (2009). Twitter sentiment classification using distant supervision. Technical report, Standford University.

[4] Alecel al. (2009).

[5] Bilal, Muhammad, et al. ((2016)). Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University-Computer and Information Sciences 28.3 , 330-344.

[6] Cole, N. L. (2017, March). What is Discourse? Retrieved from www.thoughtco.com: https://www.thoughtco.com/discourse-definition-3026070

[7] Computational Discourse. (n.d.). Retrieved from http://www3.cs.stonybrook.edu/~ychoi/cse507/slides/06-discourse.pdf

[8] Daud, Misbah, Rafiullah Khan, and Aitazaz Daud. . ((2015)). Roman Urdu opinion mining system (RUOMiS). arXiv preprint arXiv:1501.01386 .

[9] Dey, Lipika and Haque, Sk. (2009). Opinion Mining from Noisy Text Data. International Journal on Document Analysis and Recognition 12(3). , pp. 205-226.

[10] Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. (2005). Experiments on sense annotation and sense disambiguation of discourse connectives. In Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories.

[11] Forbes-Riley, Kate, Fan Zhang, and Diane Litman. (2016). Extracting PDTB discourse relations from student essays. Proc. of the SIGDIAL.

[12] Graff. (2003). English gigaword corpus.

[13] Grosz, Barbara J. and Candace L. Sidner. (1986). Attention, intentions, and the. Computational Linguistics, 12(3):175–204.

[14] Hobbs, Jerry R. (1985). On the coherence and structure of discourse. Technical Report. Stanford, C.A.: Center for the Study of Language and Information (CSLI).

[15] Joty, Shafiq, Giuseppe Carenini, and Raymond T. Ng. (2012.). A novel discriminative framework for sentence-level discourse analysis. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics,.

[16] Lüngen, Harald, et al. (2006). Discourse segmentation of german written texts. Advances in Natural Language Processing. Springer Berlin Heidelberg, (pp. 245-256).

[17] Marcu & Echihabi. (2002).

[18] Marcu, Daniel. (2000). The Theory and Practice of Discourse, Parsing and Summarisation. Cambridge, M.A.: MIT Press.

[19] Mitocariu, Elena, Daniel Alexandru Anechitei, and Dan Cristea. (2013). Comparing discourse tree structures. nternational Conference on Intelligent Text Processing and Computational Linguistics. Springer Berlin Heidelberg.

[20] Mukherjee, S. (2012). Retrieved from https://pdfs.semanticscholar.org/0399/d036564b47387ff9be1bf32261c229667c2d.pdf

[21] Nguyen, Huy, and Diane J. Litman. (2016). mproving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics. FLAIRS Conference.

[22] Polanyi et al. . (2004).

[23] Prasad et., al. (2008).

[24] Song, Wei, et al. (2015). Discourse Element Identification in Student Essays based on Global and Local Cohesion. EMNLP.

[25] Syeed Ibn Faiz and Robert E. Mercer. (n.d.). Identifying Explicit Discourse Connectives in Text. London, ON, Canada: Department of Computer Science The University of Western Ontario.

[26] Tafseer A., Saba U., Sarmad, H, Asad M., Rahila P. Farah Adeeba, Annette Hautli, Miriam Butt. (n.d.). The CLE Urdu POS Tagset.

[27] Webber, Bonnie, and Aravind Joshi. (2012). Discourse structure and computation: past, present and future. Proceedings of the ACL-2012 special workshop on rediscovering 50 years of discoveries. Association for Computational Linguistics.

[28] Wolf, Florian and Gibson, Edward. (2005). Representing discourse coherence: A corpus-based study,Computational Linguistics.

[29] Ziheng Lin, Min-Yen Kan & Hwee Tou Ng. (n.d.). Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. Department of Computer Science National University of Singapore 13 Computing Drive. Singapore.