Clustering Urdu News Using Headlines

Samia Khaliq, Waheed Iqbal, Faisal Bukhari and Kamran Malik
Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan.
E-mail: {mscsf13m014,waheed.iqbal,faisal.bukhari, kamran.malik} @pucit.edu.pk

Abstract

In this paper, we proposed and evaluated a new algorithm to automatically cluster Urdu news from different news agencies. This task is challenging as we do not have language processing libraries for Urdu language. Our experimental dataset consists of news from famous Pakistani media houses including Jang, BBC Urdu, Express, UrduPoint, and Voice of America Urdu (VOA). The proposed algorithm only uses headlines to cluster the news. News headline provide a concise summary of the news which motivates us to use it instead of using the entire news story. Our experimental evaluation shows micro and macro averages for precision 0.45 and 0.48 respectively for identifying similar news using headlines.

Keywords—*Urdu News, Clustering, Similar News, News Aggregation*

1. Introduction

Nowadays, most of the media houses publish news online to their websites and social networking sites to rapidly attract readers. Majority of the news readers are interested to read specific type of news according to their own interests. For example, politicians and businessmen are interested to remain updated with latest news especially related to politics as an abrupt change in a political scenario, not only influences the reputation of country but may also have a drastic impact on the economy of the country. Moreover, some of the media houses provide different perspective, biased or unbiased, based on a same news. Therefore, it may be helpful to read same news from different broadcasting agencies. A news aggregation tool is required to automatically aggregate news from various news sources and cluster them for the readers to read same news from different sources.

Urdu language is spoken by more than 100 million people all over the world. However, a limited research is conducted to develop Natural Language Processing (NLP) tools and APIs to process Urdu text easily. In this paper, we propose and evaluate a new algorithm to process Urdu news and cluster similar news. News articles and clustering techniques are widely used over Internet for various language, however, there is no automated service that aggregates and clusters the Urdu news from various Urdu news agencies. Finding similar or related news is beneficial for reader. As these news come from different news sources and reader can easily browse through the same news coming from different sources. Moreover many news agencies provide different perspective based on the same news. So it is important for readers to read the same news from different news agencies.

We have developed a crawler that scraps Urdu news from various Urdu news broadcasting agencies. We have an online portal (htt://www.newslink.pk) that aggregates and show the updated news from various news sources. Now, in this paper we have presented and evaluated our algorithm to automatically cluster Urdu news using only headlines. News headline provide a concise summary of the news which motivates us to use it instead of using the entire news story.

The rest of this paper is organized as follows. Section 2 provides related work of performing clustering using news. Section 3 describes our technique for clustering of Urdu news using headlines. Section 4 provides experimental evaluation and results. Section 5 concludes the paper and explains our future work.

2. Related Work

- Many clustering techniques are used by various researchers that use different criteria for clustering of News articles or reports. In [1] focuses on event centric clustering of news. They get news reports from different news sources and cluster them according to events. Their system work in online incremental environment. They have used RSS feed as a source of news report and their system clusters news reports from separate RSS feed. Their results show that their system gives much better results while using fine grained clustering technique rather coarse grained technique, which gives poor performance. They have used modified K. Means clustering technique for incremental approach of news clustering.

- Another criterion for clustering news is based on Topic. Shah and Elbahesh [2] discussed their approach for clustering web based news articles into topic wise categories. They have pointed out that search engines that search for news articles have some drawback as these search engines only search on basis of keywords and ignores the whole content of a news article. So they focused on applying text mining techniques and proposed a system that will do clustering of news articles on basis of their topic. For this purpose they have used three clustering algorithms on their dataset. K means, single link clustering algorithm and Hybrid clustering algorithm. Their results show that Hybrid algorithm outperforms other two algorithms and gives much better results than other algorithms.

Some other approaches use WordNet [3] for clustering of news articles. Similarly event centric clustering is focused on events mentioned in articles [4]. Word N-grams is also used for enhancing document clustering by the same researchers who have used WordNet [5]. Cluster centric approach is also followed to extract events from online news where already created clusters are used for extraction purpose [6].

Various news analytical tools use different techniques to classify news articles. Classification techniques are used by many researchers. A news headline has positive and negative effect on its viewers. It also contains emotions. Text of news headlines makes it positive or negative news. Santos, Ramos and Marques work on Portuguese News Headlines. They have classified news headlines as positive, negative and neutral. For this purpose they have used supervised approach and trained a classifier. This classifier classifies news headlines in above three categories. They have used two classification algorithms, Sequential Minimal Optimization (SMO). It is SVM (Support Vector Machine Method). Second classifier they used was Random Forest. These algorithms were used to recognize the features. The experiments were done on syntactic features which they explained as argument1 verb argument2 relation. Results of their experiments show that using these relations as features improves sentiment classification of news headlines [7].

Bergen and Gilpin also worked on classification of news article headlines in positive, negative and neutral news. They tried to uplift positive news headlines so it can positively affect the readers. Their goal was to develop an algorithm that distinguishes between positive and negative stories for this purpose they classify news articles headlines as positive or negative.

In [8] they have collected data for news articles from RSS feeds and sources used were Google News, CNN, Fox News, The New York Times. Bergen and Gilpin used feature extraction for both positive and negative news and two different classification algorithms were used: Naïve Bayes and Support Vector Machine also dataset was also divided in two forms, first in which only headline data was included while other include headline with text. On both datasets both classification algorithms were applied. Their results represent that Naïve Bayes classifier is better than Support Vector Machine. Naïve Bayes have accuracy around 70% while Support Vector Machine accuracy was around 68% [8].

In [9] classification of Thai news was done through structural features. In this paper news web documents from two different sources was collected. The main purpose of this paper was to formulate a simple method that extracts news articles from web collections. They have explored machine learning methods which helped in distinguishing article pages from non-article pages of web collection. After separating article pages they have compared these articles in fine grained manner so that they can identify informative structures present in the articles. In both phases of article extraction from web documents and extraction of informative structures, they have used structural features. For classification they have used three classification algorithms i.e. SVM, Naïve Bayes and C4.5. Their experimental results show that SVM works better than other two algorithms but the difference was not extraordinary.

Ramdass and Seshasai [10] also worked on classification of news articles. They have used dataset of news articles from MIT newspaper The Tech. the Tech archive requires classification of news articles into sections like News, Sports and Opinions. So the main objective of their project was to investigate and implement techniques that classify those articles into their relevant sections. They have used already classified documents so they make use of supervised classification techniques. They have split those documents for training and testing purpose. For experiments they have used different natural language feature sets and also some statistical techniques that used these feature sets. Naïve Bayes classification and Maximum Entropy Classification techniques were used for the experiments.

Their results show that news articles have two different directions one is news content and other is opinion content. The first half of their study which focused at Naïve Bayes and Maximum Entropy classifiers used the content, while the second half looked at grammatical structure. Results of both halves proved that Naïve Bayes and Maximum Entropy classifiers outperformed the results of second half.

Apart from classification, text similarity is used in document clustering. Also techniques are used for clustering of news articles. Anna Huang in [11] discussed and analyzed clustering of documents. The author has done comparison of different measures used to determine similarity of text between different documents. They have used K means algorithm for clustering and for results are compiled on seven different text datasets. Five similarity measures were evaluated by the author. Similarity measures that the author discussed are: Euclidean Distance, Cosine Similarity, Metric, Pearson Correlation Coefficient, Jacard Coefficient. Results of their experiment represent that among similarity measures Euclidean distance performs worst while Jacard and Pearson Coefficient perform better than other similarity measures and produce much better clusters.

- Thilagavathi, Anitha, and Nethra also worked on clustering of documents that is based on sentence similarity. For clustering of documents they have used fuzzy algorithm instead of hard clustering. They have mentioned that fuzzy clustering algorithm is more flexible and it allows a pattern to belong to the entire produced cluster but the degree of their membership will be different for every cluster. Fuzzy clustering algorithm works on *Expectation-Maximization Framework*. This framework helps in determining the probability of membership of a sentence in a cluster. The authors concluded that FRECCA that is a fuzzy clustering algorithm can be used for any relational problem of clustering. Their results show that fuzzy algorithm helps in avoiding the overlap and gives much better performance [12].

3. Methodology

To achieve our goal of formulating an algorithm for similar news identification following steps are used:

A. Data Gathering:

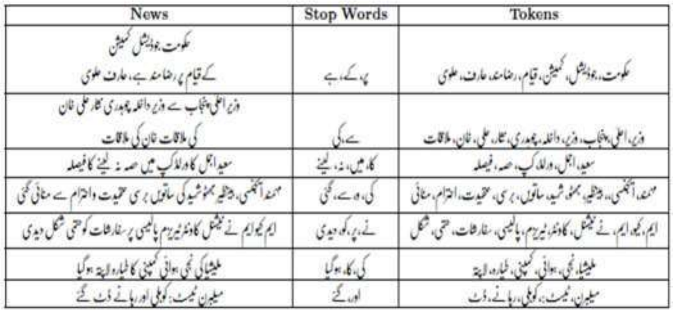
We have developed a crawler web crawler which gathers headlines from different media houses websites. We have targeted renowned news channels of Pakistan. The crawler is capable to scrap news headline, news image, news date and time, and news story. Currently our crawler is scraping news from the following media houses websites:

- BBC Urdu
- ARY News
- Nawa eWaqt
- Express News
- Jang
- GEO News
- UrduPoint

B. Pre-Processing:

In pre-processing phase we cleaned the news headlines by removing stop words and identify tokens for each news headline. Table I shows the pre-processing of few news headlines. The Table provides news, stop words, and tokens identified for some sample headlines.

Table I: Pre-processing news headlines to identify tokens by removing stop words.



News	Stop Words	Tokens
حکومت، جوڈیشل کمیشن کے قیام پر رضامند ہے، عارف علوی	پر، کے، ہے	حکومت، جوڈیشل، کمیشن، قیام، رضامند، عارف، علوی
وزیر اعلیٰ پنجاب سے وزیر داخلہ چوہدری نثار علی خان کی ملاقات خان کی ملاقات	سے، کی	وزیر، اعلیٰ پنجاب، وزیر، داخلہ، چوہدری، نثار، علی، خان، ملاقات
سعید اجمل کا ورلڈکپ میں حصہ نہ لینے کا فیصلہ	کا، میں، نہ، لینے	سعید، اجمل، ورلڈکپ، حصہ، فیصلہ
مہند آنگلس، ویٹیلی ہمنوشیو کی ساتوں بری حقیت و الزام سے منافی گئی	کی، و، سے، گئی	مہند، آنگلس، ویٹیلی، ہمنوشیو، ساتوں، ب ری، حقیت، الزام، منافی
ایم کیو ایم نے نیشنل کاونٹر ٹیررزم پالیسی پر سفارشات کو حتمی شکل دیدی	نے، پر، کو، دیدی	ایم، کیو، ایم، نے نیشنل، کاونٹر ٹیررزم، پالیسی، سفارشات، حتمی، شکل
ملیشیا کی نجی ہوائی کمپنی کا طیارہ لاپتہ ہوگیا	کی، کا، ہوگیا	ملیشیا، نجی، ہوائی، کمپنی، طیارہ، لاپتہ
میلبرن ٹیسٹ: کوہلی اور رہانے ڈٹ گئے	اور، گئے	میلبرن، ٹیسٹ، کوہلی، رہانے، ڈٹ

C. News Clustering Algorithm:

We have design an algorithm to identify similar news. Following are the notations and formulas uses to design the algorithm:

- n_i – news i
- n_j – news j
- $S_{i,j}$ – similarity score between news i and news j
- tl_i - token list of news i
- tl_j – token list of news j
- st_i – size of token list
- st_{avg} – average size of token lists of news i and j
- $m_{i,j}$ – count of similar tokens of news i and news j
- t – a constant threshold value

Where;

$$stavg= \frac{sti+ stj}{2} \qquad (1)$$

$$S_{i;j} = \frac{m_{i,j}}{st_{avg}} \qquad (2)$$

The threshold variable is used by the algorithm to identify similar news based on matching number of tokens between given two news headlines. For example, if the calculated similarity score $S_{i;j}$ is greater than or equal to threshold value then both news headlines will be considered as similar. The main part of the algorithm is to identify the list of similar news on a given news headlines. We model this algorithm as a function named getRelatedNewsList() and explained in Algorithm 1.

4. Experimental Evaluation

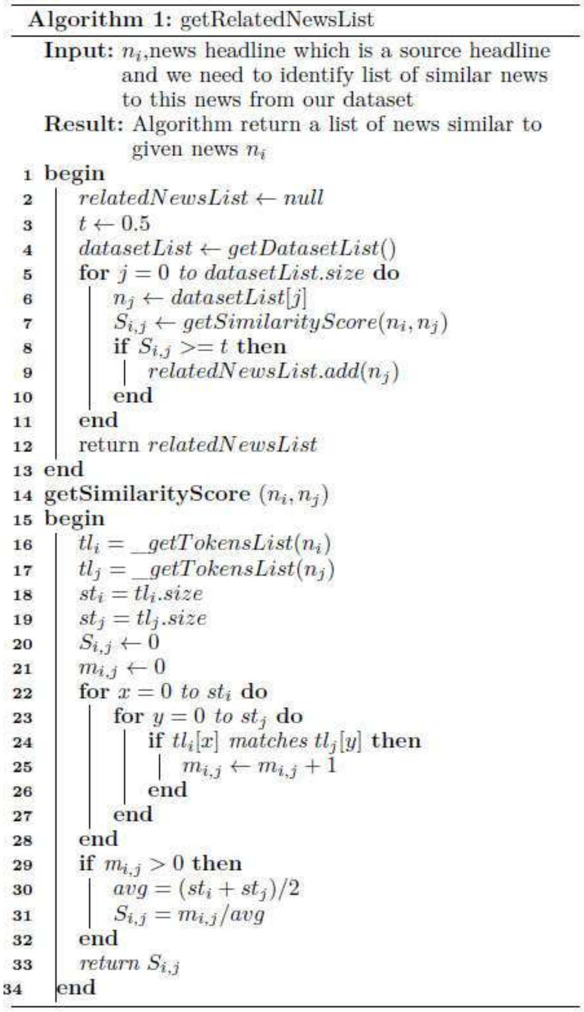
To evaluate results of clustered news we used a dataset consisting on 500 news headlines. These news headlines are distributed on the following five categories equally:

- International
- National
- Health
- Business
- Entertainment

First we run our algorithm on these categories which give clusters of related news headlines. Now these clusters are compared with the ground truth.

A. Formatting Ground Truth

To compare results of system generated clusters we have defined ground truth for each category. Ground truth is devised manually. We have given each category news headlines to three persons and ask them to mark related news. Then we have compared related news marked by all participants and select those clusters of related news that are common or marked by at least two participants. In this way we get ground truth of clusters for each category.



```
Algorithm 1: getRelatedNewsList
Input: $n_i$,news headline which is a source headline
       and we need to identify list of similar news
       to this news from our dataset
Result: Algorithm return a list of news similar to
        given news $n_i$
begin
    $relatedNewsList \leftarrow null$
    $t \leftarrow 0.5$
    $datasetList \leftarrow getDatasetList()$
    for $j = 0$ to $datasetList.size$ do
        $n_j \leftarrow datasetList[j]$
        $S_{i,j} \leftarrow getSimilarityScore(n_i, n_j)$
        if $S_{i,j} >= t$ then
            $relatedNewsList.add(n_j)$
        end
    end
    return $relatedNewsList$
end
getSimilarityScore $(n_i, n_j)$
begin
    $tl_i = \_getTokensList(n_i)$
    $tl_j = \_getTokensList(n_j)$
    $st_i = tl_i.size$
    $st_j = tl_j.size$
    $S_{i,j} \leftarrow 0$
    $m_{i,j} \leftarrow 0$
    for $x = 0$ to $st_i$ do
        for $y = 0$ to $st_j$ do
            if $tl_i[x]$ matches $tl_j[y]$ then
                $m_{i,j} \leftarrow m_{i,j} + 1$
            end
        end
    end
    if $m_{i,j} > 0$ then
        $avg = (st_i + st_j)/2$
        $S_{i,j} = m_{i,j}/avg$
    end
    return $S_{i,j}$
end
```

B. Evaluation Criteria:

After getting ground truth for each category we have evaluated system generated clusters and computed different evaluation measures like precision, recall, F measures. A confusion matrix for category "National" is shown in Table II. Similarly we computed confusion matrix for each category before computing the evaluation metrics.

Table II: Confusion Matrix for Category National.

		Predicted	
		UNC	CN
Actual	UNC	TN =91	FP =1
	CN	FN = 3	TP = 5

Following are the symbols used in the equations to calculate different evaluation metrics:

- TN: True negative are the number of news that are not included in any cluster by ground truth and by our system
- FN: False Negative number of news that our system did not find any cluster but in ground truth these news are included in clusters.
- FP: False Positive number of news that our system includes in cluster but in ground truth these news are not clustered
- TP: True Positive number of news clustered by both ground truth and by system.
- NP: News Population is total number of news headlines in a category.
- TFN: Total False Negative number of news that our system did not find any cluster but in ground truth these news are included in clusters.
- TFP: Total False Positive number of news that our system includes in cluster but in ground truth these news are not clustered
- TTP: Total True Positive number of news clustered by both ground truth and by system.
- UNC: Number of news that are not in any cluster
- CN: Number of news that are part of specific cluster.

For each category we computed Precision (P), Recall(R) and F1 Measure (F1) from this confusion matrix using the following formulas:

$$P = \frac{\text{TP}}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F\,1 = \frac{2\text{PR}}{P + R} \tag{5}$$

Overall Macro average Precision (M_{acp}), Macro average Recall (M_{acR}), Macro average F-Measure (M_{acF1}), Micro average Precision (M_{icp}), Micro average Recall (M_{icR}) and Micro average F1 Measure (M_{icF1}) of all categories are calculated using the following formulas:

$$M_{acp} = \frac{1}{n}\sum_{i=1}^{n} P(C_i) \tag{6}$$

Where $P(C_i)$ represents the Precision for specific i^{th} category.

$$M_{acR} = \frac{1}{n}\sum_{i=1}^{n} R(C_i) \tag{7}$$

Where $R(C_i)$ represents the Recall for specific i^{th} category.

$$M_{acF1} = \frac{1}{n}\sum_{i=1}^{n} F1(C_i) \tag{8}$$

Where $F\,1(C_i)$ represents the F-Measure for i^{th} category.

$$M_{icp} = \frac{TTP}{TTP+TFP} \tag{9}$$

$$M_{icR} = \frac{TTP}{TTP+TFN} \tag{10}$$

$$M_{icF1} = \frac{2*M_{icp}*M_{icR}}{M_{\ icp}+M_{icR}} \tag{11}$$

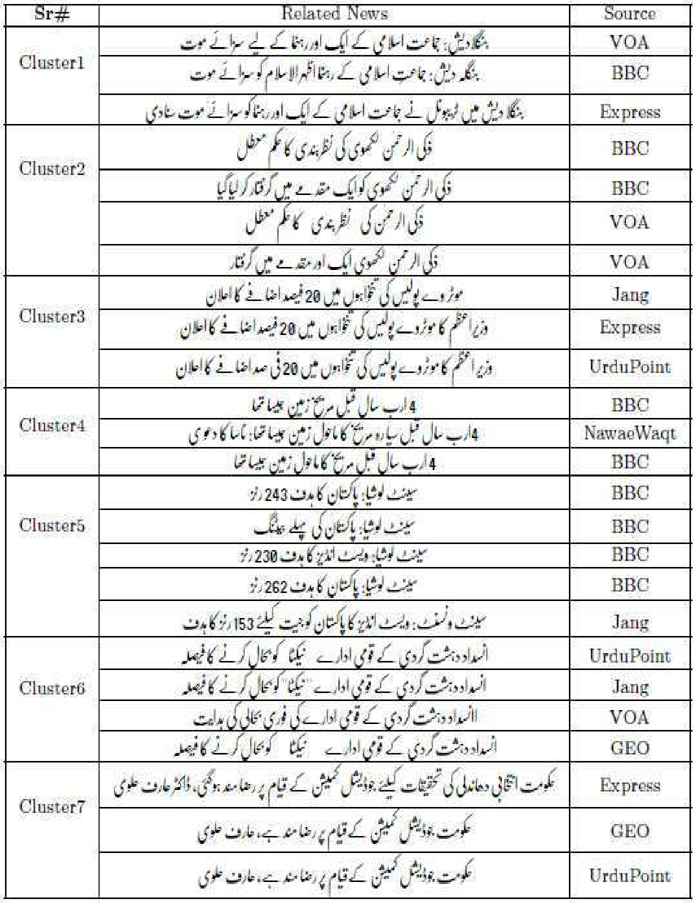
C. Results:

Table III shows category wise Precision, Recall, and F-Measure. We also show micro and macro averages for each of these measures. The highest Precision (0.84) is achieved in National category, however, the lowest Precision (0.26) is found in Business category. The micro average for Precision is obtained is 0.45, however, the macro average for Precision is obtained 0.48. The result looks far better than a random probability of (0.2) for Precision. Therefore, the proposed algorithm is effective to cluster similar news. We show some sample clusters identified by our proposed algorithm in Table IV.

Table III: Category wise Precision, Recall F-Measure

Category	Precision	Recall	F-Measure
International	0.62	0.5	0.55
National	0.83	0.63	0.71
Health	0.43	0.5	0.46
Business	0.26	0.35	0.3
Entertainment	0.29	0.33	0.31
Micro Average	0.45	0.46	0.46
Macro Average	0.48	0.46	0.47

Table IV: Clustering Results



Sr#	Related News	Source
Cluster1	بنگلادیش: جماعت اسلامی کے ایک اور رہنما کے لیے سزائے موت	VOA
	بنگلہ دیش: جماعت اسلامی کے رہنما اظہر الاسلام کو سزائے موت	BBC
	بنگلادیش میں ٹریبونل نے جماعت اسلامی کے ایک اور رہنما کو سزائے موت سنادی	Express
Cluster2	ذکی الرحمن لکھوی کی نظربندی کا حکم معطل	BBC
	ذکی الرحمن لکھوی کو ایک مقدمے میں گرفتار کر لیا گیا	BBC
	ذکی الرحمن کی نظربندی کا حکم معطل	VOA
	ذکی الرحمن لکھوی ایک اور مقدمے میں گرفتار	VOA
Cluster3	موٹروے پولیس کی تنخواہوں میں 20 فیصد اضافے کا اعلان	Jang
	وزیراعظم کا موٹروے پولیس کی تنخواہوں میں 20 فیصد اضافے کا اعلان	Express
	وزیراعظم کا موٹروے پولیس کی تنخواہوں میں 20 فی صد اضافے کا اعلان	UrduPoint
Cluster4	4 ارب سال قبل مریخ زمین جیسا تھا	BBC
	4 ارب سال قبل سیارہ مریخ کا ماحول زمین جیسا تھا: ناسا کا دعوی	NawaeWaqt
	4 ارب سال قبل مریخ کا ماحول زمین جیسا تھا	BBC
Cluster5	سینٹ لوشیا: پاکستان کا ہدف 243 رنز	BBC
	سینٹ لوشیا: پاکستان کی پہلے بیٹنگ	BBC
	سینٹ لوشیا: ویسٹ انڈیز کا ہدف 230 رنز	BBC
	سینٹ لوشیا: پاکستان کا ہدف 262 رنز	BBC
	سینٹ ونسنٹ: ویسٹ انڈیز کا پاکستان کو جیت کیلئے 153 رنز کا ہدف	Jang
Cluster6	انسداد دہشت گردی کے قومی ادارے نیکٹا کو بحال کرنے کا فیصلہ	UrduPoint
	انسداد دہشت گردی کے قومی ادارے "نیکٹا" کو بحال کرنے کا فیصلہ	Jang
	انسداد دہشت گردی کے قومی ادارے کی فوری بحالی کی ہدایت	VOA
	انسداد دہشت گردی کے قومی ادارے نیکٹا کو بحال کرنے کا فیصلہ	GEO
Cluster7	حکومت انتخابی دھاندلی کی تحقیقات کیلئے جوڈیشل کمیشن کے قیام پر رضامند ہوگئی، ڈاکٹر عارف علوی	Express
	حکومت جوڈیشل کمیشن کے قیام پر رضامند ہے، عارف علوی	GEO
	حکومت جوڈیشل کمیشن کے قیام پر رضامند ہے، عارف علوی	UrduPoint

5. Conclusion and Future Work

Urdu language readers are presented all over the world. One of the popular things for these readers is to read Urdu news from the Internet. In this paper, we presented and evaluated an algorithm to automatically cluster similar news from various Urdu news media houses. Our experimental evaluation shows an average micro average for Precision measure is 0.45 and the macro average for Precision is 0.48. We believe that the work will help us to provide a tool for Urdu news readers to read same news from different broadcasting services to understand different perspective on the same news.

Currently, we are integrating this work with our on line portal (newslink.pk). We are also looking to improve the processing time of the algorithm by using Apache Hadoop.

6. References

[1] J.Azzopardi and C.Staff, "Incremental clustering of news reports," *Algorithms*, vol. 5, no. 3, pp. 364–378, 2012.

[2] N. A. Shah and E. M. ElBahesh, "Topic based clustering of news articles," in *Proceedings of the 42Nd Annual Southeast Regional Conference*, ser. ACM-SE 42. New York, NY, USA

[3] C. Bouras and V. Tsogkas, "W-kmeans: Clustering news articles using WordNet," in *14th International Conference on Knowledge Based and Intelligent Information andEngineering Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 379–388.

[4] J. Borglund, "Event centric clustering of news articles," Uppsala University, Department of Information Technology, Tech. Rep. 13 072, 2013.

[5] C. Bouras and V. Tsogkas, "Enhancing news articles clustering using word n-grams," in *2nd International Conferenceon DataManagement Technologies and Applications*, 2013.

[6] J. Piskorski, H. Tanev, M Atkinson and E. Van Der Goot, "Cluster centric approach to news event extraction," *in Proceedings of 2008 Conference on New Trends in Multimedia and Network Information Systems*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2008, pp.276–290. http://dl.acm.org/citation.cfm?id=1565754.1565782

[7] C. Ramos and N. C. Marques, "Sentiment classification of Portuguese news headlines" in International Journal of Software Engineering and its Applications, vol. 9, Portugal,2015, pp. 9-18. http://dx.doi.org/10.14257/ijseia.2015.9.9.02

[8] K. Bergen and G. Leilani, "Negative news no more: Classifying news article headlines," Stanford University, USA, Tech. Rep., 2012.

[9] S. Tongchim, S. Virach, and H. Isahara, "Classification of news web documents based on structural features," in *Advances in Natural Language Processing*, vol. 4139, 2006, pp. 153–160

[10] D. Ramdass and S. Shreyes, "Document classification for newspaper articles," 2009.

[11] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*, 2008, pp. 49–56.

[12] K. G.Thilagavathi, J.Anitha, "Sentence similarity based document clustering using fuzzy algorithm," *International Journal of Advance Foundation and Research in Computer*, vol. 1, 2014.