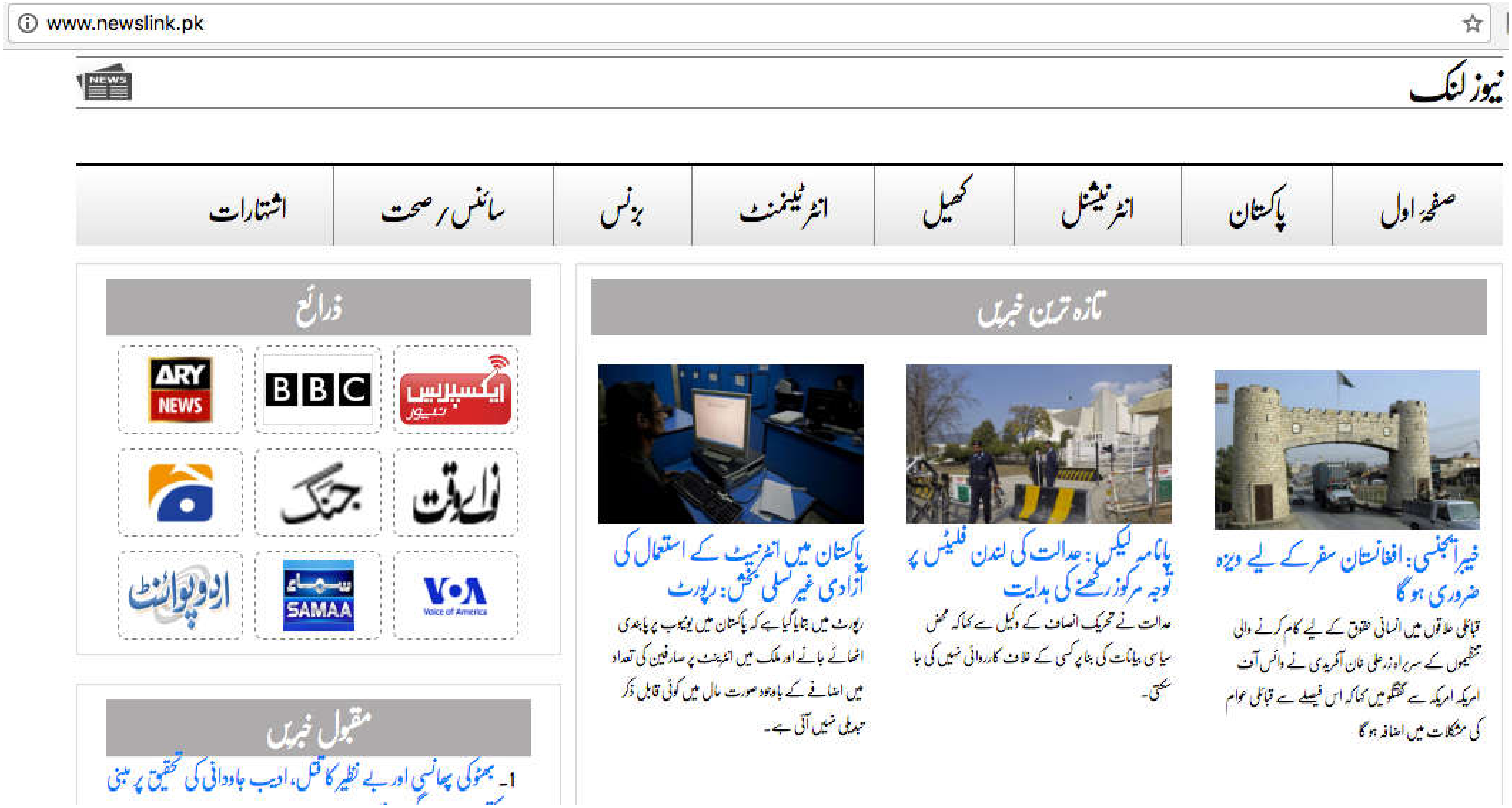
# Clustering Urdu News Using Headlines

Samia Khaliq, **Waheed Iqbal,** Faisal Bukhari, Kamran Malik
PUCIT, University of the Punjab,
Lahore, Pakistan

# Motivation


www.newslink.pk
NEWS
نیوز لنک
صفحہ اول
پاکستان
انٹرنیشنل
کھیل
انٹرٹینمنٹ
بزنس
سائنس / صحت
اشتہارات
تازہ ترین خبریں
خیبر ایجنسی: افغانستان سفر کے لیے ویزہ ضروری ہو گا
قبائلی علاقوں میں انسانی حقوق کے لیے کام کرنے والی تنظیموں کے سربراہ زر علی خان آفریدی نے وائس آف امریکہ سے گفتگو میں کہا کہ اس فیصلے سے قبائلی عوام کی مشکلات میں اضافہ ہو گا
پاناما لیکس: عدالت کی لندن فلیٹس پر توجہ مرکوز رکھنے کی ہدایت
عدالت نے تحریک انصاف کے وکیل سے کہا کہ محض سیاسی بیانات کی بنا پر کسی کے خلاف کارروائی نہیں کی جا سکتی۔
پاکستان میں انٹرنیٹ کے استعمال کی آزادی غیر تسلی بخش: رپورٹ
رپورٹ میں بتایا گیا ہے کہ پاکستان میں یوٹیوب پر پابندی اٹھائے جانے اور ملک میں انٹرنیٹ پر صارفین کی تعداد میں اضافے کے باوجود صورت حال میں کوئی قابل ذکر تبدیلی نہیں آئی ہے۔
ذرائع
ARY NEWS
BBC
ایکسپریس نیوز
جنگ
نوائے وقت
اردو پوائنٹ
SAMAA
Voice of America
مقبول خبریں
1۔ بھٹو کی پھانسی اور بے نظیر کا قتل، ادیب جاودانی کی تحقیق پر مبنی

# Motivation (Cont.)

صفحۂ اول | پاکستان | انٹرنیشنل | کھیل | انٹرٹینمنٹ | بزنس | سائنس / صحت | اشتہارات

اس جیسی خبریں

خبر کی تفصیل

ہنڈی حوالہ کے کاروبار میں ملوث ملزم گرفتار

لاہور (خبرنگار) ایف آئی اے کے اسسٹنٹ ڈائریکٹر احسان اللہ خان نے شاہ عالمی میں فوارہ چوک قریب مارکیٹ میں

# Introduction

- We are actively scraping Urdu news from news agencies including Jang, Geo, Express, Nawa-e-Waqat etc.
- Currently we have around 0.5 million news scrapped from different news websites.
- Our aim is to develop a portal that can provide Urdu news analytics.
- In this work, we presented our algorithm to automatically cluster similar Urdu news.

# News Scrapping

- We use Selenium to scrap the news websites.
- We configured a cron job to run the scrapper after every 4 hours.
- We scrap around 700 to 1200 news every day.

# Pre-Processing

News headlines are cleaned in this step by removing stop words and identifying tokens.

| News | Stop Words | Tokens |
|---|---|---|
| حکومت .جوڈیشل کمیشن<br>کے قیام پر رضا مند ہے، عارف علوی | پر، کے، ہے | حکومت، جوڈیشل، کمیشن، قیام، رضامند، عارف، علوی |
| وزیر اعلیٰ پنجاب سے وزیر داخلہ چوہدری نثار علی خان<br>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، ورلڈ، کپ، حصہ، فیصلہ |
| مہمند ایجنسی، بینظیر بھٹو شہید کی ساتویں برسی عقیدت واحترام سے منائی گئی | کی، و، سے، گئی | مہمند، ایجنسی،، بینظیر، بھٹو، شہید، ساتویں، برسی، عقیدت، احترام، منائی |
| ایم کیو ایم نے نیشنل کاونٹر ٹیریزم پالیسی پر سفارشات کو حتمی شکل دیدی | نے، پر، کو، دیدی | ایم، کیو، ایم، نے نیشنل، کاونٹر، ٹیریزم، 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 |

# New Similarity

Given two news *i* and *j* we need to identify similarity between them:

$$st_{avg} = \frac{st_i + st_j}{2}$$

$$S_{i,j} = \frac{m_{i,j}}{st_{avg}}$$

$st_i$ - size of tokens list of news i

$st_j$ - size of tokens list of news j

$st_{avg}$ - average size of token lists of news i and j

$m_{i,j}$ - count of similar tokens of news i and news j

**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 ← null
    t ← 0.5
    datasetList ← getDatasetList()
    for j = 0 to datasetList.size do
        n_j ← datasetList[j]
        S_i,j ← 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 ← 0
    m_i,j ← 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 ← 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
```

# Ground Truth

- We randomly selected 100 news from each of the following categories:
    - International
    - National
    - Health
    - Business
    - Entertainment

- Each news category is given to three different persons who marked related news manually. Then we identified the common clusters as ground truth in each of the category.

# Evaluation Criteria

- We use the same 500 news to run our news similarity algorithm.
- The algorithm gives us cluster of news. Then we computer Confusion Matrix by comparing it with our Ground Truth.
- Now system generated clusters are compared with ground truth clusters and for each category Confusion Matrix is created. Confusion Matrix for category National is given below:

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

# Evaluation Measures

- F-measure
- Precision
- Recall

# Evaluation Results

| 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 |

# Results

| Sr# | Related News | Source |
|---|---|---|
| Cluster1 | بنگلادیش: جماعت اسلامی کے ایک اور رہنما کے لیے سزائے موت | VOA |
| | بنگلہ دیش: جماعتِ 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 الرحمٰن لکھوی کی نظربندی کا حکم معطل | BBC |
| | ذکی الرحمٰن لکھوی کو ایک مقدمے میں گرفتار کر لیا گیا | BBC |
| | ذکی الرحمٰن کی نظربندی کا حکم معطل | VOA |
| | ذکی الرحمٰن 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 |
| | اانسداد دہشت گردی کے قومی ادارے کی فوری بحالی کی ہدایت | VOA |
| | انسداد دہشت گردی کے قومی ادارے نیکٹا کو بحال کرنے کا فیصلہ | GEO |

# Results (Cont.)

# Results (Cont.)

صفحۂ اول | پاکستان | انٹرنیشنل | کھیل | انٹرٹینمنٹ | بزنس | سائنس رصحت | اشتہارات

## خبر کی تفصیل

### پنجاب میں 1200 سے زائد افراد سموگ سے متاثر ہو گئے

لاہور (اُردو پوائنٹ اخبارتازہ ترین ۔ 05 نومبر2016ء ) لاہور میں سموگ مکمل طور پر ختم نہ ہو سکی ، صبح کے اوقات میں سموگ میں کمی اور شام میں اس کی شدت بڑھ جاتی ہے ، سموگ سے بارہ سو سے زائد افراد متاثر ہو کر ہسپتال پہنچ گئے۔ دوسری طرف پنجاب کے مختلف علاقے دھند کی لپیٹ میں رہے ، موٹر وے کئی مقامات پر بند رہی۔ اس صورتحال کے پیش نظر ماہرین نے شہریوں کو احتیاط برتنے کی اپیل ہے۔ ماہرین کا کہنا ہے کہ شہری ماسک

## اس جیسی خبریں

1۔ پنجاب میں سموگ کا راج برقرار
2۔ پنجاب میں دس افراد کو پھانسی دے دی گئی
3۔ پنجاب میں مزید 24 افراد میں ڈینگی وائرس کی تصدیق
4۔ پنجاب میں ڈینگی وائرس سے متاثرہ افراد کی تعداد 54 ہو گئی
5۔ پنجاب میں مزید69 افراد میں ڈینگی وائرس کی تصدیق
6۔ سندھ،پنجاب انتخابات میں 50 افراد زخمی
7۔ انسانی اسمگلنگ میں ملوث مزید چھ افراد پنجاب سے گرفتار
8۔ نابینا افراد پنجاب حکومت کیخلاف سڑکوں پر
9۔ پنجاب میں فورسز کا آپریشن، مزید 250 افراد گرفتار
10۔ پنجاب میں سرچ آپریشن؛ 200 افراد زیر حراست

# Conclusion and Future Work

- We have an algorithm in place to integrate with newslink.pk to identify similar news automatically.
- Experimental evaluation shows an average micro average for Precision measure is 0.45 and the macro average for Precision is 0.48.
- In future, we intend to use LDA to identify similar news and compare it with our work.

# THANKS!