# An Integrated Text Mining Platform for Monitoring of Persian News Agencies

**Mohammad Taghipour[1], Foad Aboutorabi[2], Vahid Zarrabi[3], Habibollah Asghari[4]**
[1,2,3,4] ICT Research Institute, Academic Center for Education Culture and Research
Tehran, Iran
{taghipour, foad.aboutorabi, vahid.zarrabi, habib.asghari}@ictrc.ac.ir

**Abstract**

There is an increasing trend in design and development of standalone tools and techniques for natural language processing (NLP) purposes. In recent years, the news agencies have focused on extracting knowledge from a huge amount of pile text news from various media. However, little work has been done to develop a unified platform for mining and monitoring the news agencies in Persian. In this paper, we present an integrated platform for monitoring the Persian news agencies. This platform consists of four main blocks including the web/social media crawler, feature extraction, impact analysis and the visualizer. Various open source tools and techniques have been employed in order to design and implement each of the mentioned segments. The final platform has been deployed in one of the most influential Iranian news agencies as a decision support system for comparing the position and rank of the news agency with respect to the other competitors.

**Keywords:** Persian Language, News Monitoring, Near Duplicate Detection

## 1. Introduction

In today's world, the news agencies play an important role in shaping the mindset of people and their perception as well as the image of the governments in international media coverage. Surviving and having a strong presence in today's competitive world requires employing of strong tools and techniques and news agencies are not exempted from this principle. Natural Language Processing (NLP) methods can greatly help to improve the impact and performance of the news agencies. A news monitoring system tries to analyze and discover hidden patterns in data and present them in the form of key performance indices (KPI's) and also to visualize them for managerial purposes.

In this paper, we focus on a Persian news monitoring system for extracting knowledge from news agencies to properly demonstrate their position in a media ecosystem in the country. Tackling the problematic characteristics of Persian as an Arabic script-based language is one of the most challenging tasks of this research. Moreover, Persian is a low-resource language with little tools and data available in digital format. So, in order to overcome these problems and to improve the quality of the platform, some custom tools and corpora were also developed and implemented.

To evaluate the power and influence of each news agency, a collection of standard measures alongside several custom features were defined and implemented. Furthermore, various methods and techniques were developed in order to demonstrate the unusual changes in the news by some reporters and agencies. A comprehensive set of visualizations were also presented based on the data and defined measures.

The paper is organized as follows: section 2 describes the related work. Section 3 comes with system design and architecture. In section 4 we describe the implementation process. Conclusion and recommendations for future works are presented in the final section.

## 2. Related Work

For many languages around the world, various tools and techniques are available for NLP related systems and applications, but little work have been done for monitoring of news agencies in Persian. In a research accomplished by Volkovich et al., (2016), they have proposed a novel method for analyzing Arabic media using some quantitative characteristics. Their methods try to demonstrate the ways in which important social events can be recognized by analyzing two well-known Arabic daily newspapers Al-Ahraam and Al-Akhbaar. (Volkovich et al., 2016).

In a framework proposed by (Martins et al., 2016) they mine the behavior of the crowds for temporal signals. This new time aware ranking method integrates lexical, domain and temporal evidences from multiple Web sources to rank microblog posts. Their system explores the signals from Wikipedia, news articles, and Twitter feedback to estimate the temporal relevance of search topics.

Text mining technics have been used in (Nassirtoussi et al., 2015) to predict FOREX market based on news-headlines. They proposed a novel approach to predict intraday directional-movements of a currency-pair in the foreign exchange market based on the text of breaking financial news-headlines. They have addressed accessing the fundamental data hidden in unstructured text of news as a challenge in a specific context by bringing together natural language processing and statistical pattern recognition as well as sentiment analysis to propose a system that predicts directional-movement of a currency-pair in the foreign exchange market based on the words used in adjacent news-headlines in the past few hours.

## 3. System Design and Architecture

The architecture of the media monitoring system along with the main blocks of the system is described in detail in the following subsections.

### 3.1 Web/Social media Crawler

**Web Crawler:** One of the most challenging parts of any text processing platform is the way that the data is collected from the web and stored. In this research, we have used web crawlers for gathering news data through the web. The available open source crawlers have various features and characteristics that make the benchmarking process essential for choosing the most appropriate one.

It is critical for a crawler used in a news monitoring system to be scalable and robust. Moreover, Quality, Freshness and Coverage are the other important features that should be covered by a crawler. Table 1 shows a comparison between the features of some available popular Web crawlers.

| Feature / Platform | language | Operating System | License | Parallel |
|---|---|---|---|---|
| Scrapy | python | Linux/Mac OSX/Windows | BSD License | Yes |
| Apache Nutch | Java | Cross-platform | Apache License 2.0 | Yes |
| Heritrix | Java | Linux/Unix-like/Windows Unsupported | Apache License | Yes |
| Web-Sphinix | Java | Windows/Mac/ Linux/Android/ IOS | Apache Software License | Yes |

Table 1: Comparison of various open-source web crawlers (Yadav and Goyal, 2015)

We used Apache Nutch™ web crawler toolkit[1] for crawling the news because it satisfies the mentioned criteria. Apache Nutch™ also has some other features makes it outstanding among the other crawlers such as compatibility with Apache Hadoop™. It is also compatible with other Apache frameworks like Tika and Solr.

Using Alexa[2] site ranking system and consultation with media experts, 200 top ranked major news agencies were hand-picked for the project. It's also worth mentioning that the number of news crawled are estimated around 100,000 records per day. This data is stored in database and used for visualization.

Since some news agencies change the contents of the news after publishing, so all the crawled news are re-crawled again after 24 hours of their release time in order to find the unusual changes after the release.

A custom python-based RSS crawler was designed and developed for gathering the recent news. This RSS crawler has trained and scheduled using one of the well-known Reinforcement Learning (RL) algorithms. Q-Learning method was introduced by Watkins (1989). This model consists of states and set of actions per state, where each agency was considered as a state. The agent checks the RSS feed of the news agency while visiting the corresponding state. There are some times like at midnight, which the crawler should check the RSS feed less frequently, therefore, we considered a special state to address these cases, and when the agent visits it, the crawler does nothing. Each transition between states $i$ and $j$ was considered as an action between them. Q-Learning finds an optimal action-selection policy for optimum crawling. The number of news crawled in each iteration, the time interval between the iterations, and the capacity of each agency RSS feed are the features exploited for training this crawler. The aim of the method is to minimize the number of RSS feed check, while maximizing the number of crawled news in each check as well.

**Social media crawler**: Telegram™ is a well-known instant messaging service which is widely used in Iran. Unofficial statistics show that more than 40 million people use the services provided by this social media messaging service. Thousands of official and unofficial active channels are providing news to the masses. Most of the well-known agencies are also providing news through their own channels in Telegram. The above-mentioned points depict that this social media has a great value for monitoring and analyzing the media. It's also worth mentioning that the Telegram-API provides us some parameters such as view count for each post and participant count for each channel which are invaluable information to methods like Hot Topic Detection and so on. We developed a custom Python-based crawler to fetch the posts of about 5000 Telegram channels.

For extracting the Telegram data, two distinct custom crawlers were developed. The first one constantly checks the Telegram channel feeds and stores new posts in raw repository. The second crawler scheduled to check the recent posts constantly and update the view count of each post.

### 3.2 Feature extraction

**HTML/JSON Parser:** As the raw news data fetched by the crawlers couldn't be used directly, several custom parsers are developed to extract the features for next blocks of the project. We also developed a special parser for 14 more influential agencies to extract more detailed features, while a fast one is implemented for the others. The most important features which are extracted from the news in this step includes: the title, URL, release date and time, body of news, news category, report information and so on. The Telegram-API provides the data in neat format, so we easily extracted and parsed the required data.

**Preprocessing**: As the text extracted from the Web are written by different authors with various types of writing and encoding styles, a preprocessing step is required before applying any NLP task. The text extracted by the parser is fed into the Parsivar preprocessing tool (Mohtaj, et al, 2018) and then stored in a news repository. After applying a normalizer, each word in the text is marked

[1] http://nutch.apache.org/
[2] https://www.alexa.com/

corresponding to its particular Part of Speech (PoS) by its PoS tagger. Furthermore, named entities are labeled using Stanford Named Entity Recognizer (Finkel, J. R., et al., 2005). For improving the results of the NER algorithm, a customized Persian NER algorithm was also employed (Zafarian, A., et al., 2015). In the NER module, we extract the Names, Places and Time from the body of the news.

**Similarity detection:** Near-duplicate detection in news is the primary step to calculate the other features. We need a fast method to find similar news and to compute the news and agency characteristics, assuming that a great deal of news is released daily. Min-hash is a popular algorithm for similarity search, especially for high dimensional data in which it converts large sets into short signatures while preserving similarity (Theobald, et al 2008). We used Min-hash to extract short signatures and Jaccard similarity for computing similarity between signatures. Finally, we clustered all pairs of news which are highly similar.

**Automatic news labeling:** All of the news agencies are hand-labeled in some categories. Since the default categories used by each agency are different from the other agencies, so we should generate a unified class of category and label all the news according to it. We select a base category from one of the top ranked news agencies. Using the SVM method, a classifier was trained and employed for mapping the news to our standard category.

**News topic detection:** In order to get a deeper insight from the data, a multi-level clustering engine based on Latent Dirichlet Allocation (LDA) method was implemented on the crawled data (Blei et al. 2003). LDA is one of the most powerful methods for modeling topics of the documents which is able to properly extract the topics and provide a deep insight to the users. As the original LDA method is inherently not incremental, so we implemented another version of LDA which is capable of dealing with news streams (Hoffman, et al., 2010).

### 3.3 Impact analysis

Using the data and features collected from the previous steps, several quantitative characteristics or indicators are derived and calculated in this step. These characteristics are especially useful when the users want to know about the influence and penetration of the news, agencies and reporters. In order to describe the indicators, some definitions should be clarified as follow:

*Media:* The web based infrastructure of news platforms including news agencies, news websites, newspapers and social media.

*Target News*: The current news that the other indicators are calculated based on it.

*Target Media:* The current media that released the Target News.

*Release time:* The time that the Target News is released by an agency.

*Release Chain*: All the similar news copied and released by the other agencies are put in a time-line which is called the Release Chain.

*Starter Media*: A news agency or media that is the starter of a release chain

*Starter News*: A news that is the starter of a release chain

Table 2 illustrates some of the indicators used in the system.

| Indicator | Calculation Method |
|---|---|
| Absolute delay | The time interval between the release of target news and the release of the starter news in the release chain |
| Relative delay | The average time interval between the release of target news and the release of all previous news in the release chain |
| Release rank | The rank of the current agency in the release chain (number of released news before the target news in the release chain) |
| News penetration | The total number of news in the release chain |
| Agency penetration | The total number of agencies in the release chain that released the target news after the current agency |
| Similarity Rate | The similarity percentage between the target news and the starter news. This parameter is also referred to as impact |
| Change factor | The degree to which a news changes with respect to its first release |
| Rate of Micro News | Micro News are the news with no significant information because of their short length, which sometimes are just headlines. As these news may be completed after the release time, they are considered as a negative indicator in media evaluation. |

Table 2: News characteristics and their estimation methods

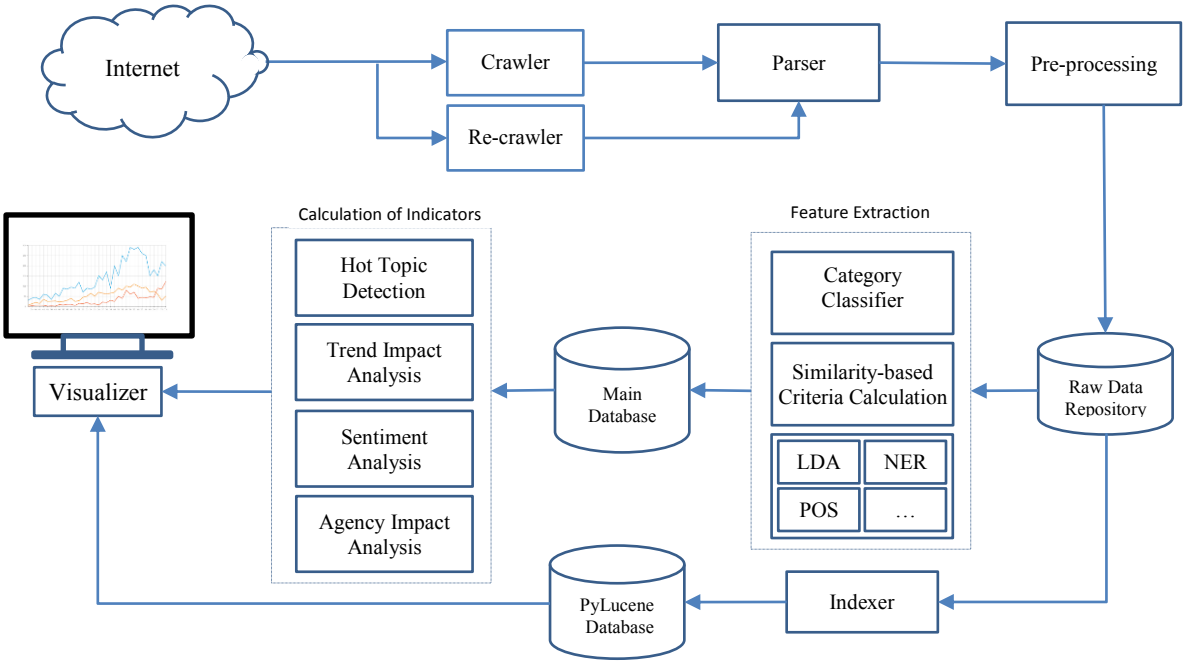


Figure 1: The complete system design architecture

### 3.6 Visualizer

For visualizing the indicators and parameters extracted from the previous steps, we constructed a web-based platform. Since most of the libraries used in this project were based on python APIs, so we used Django as a robust and powerful web-framework. Moreover, in the client side, several JavaScript frameworks such as JqueryUI[3], Telerik Kendo UI® Core[4], Gephi[5] and other frameworks were used. We divided the visualizations into three categories of News, Agencies, and Reporters.

In the 'News' section, the user can thoroughly search in the crawled news and observe the above-mentioned features and indicators. In the 'Agencies' section, the reports were divided into two categories of time-dependent and non-time-dependent charts. The non-time-dependent charts generally report the number of news published by each agency in a specific time frame. They also report the numbers and figures related to the modified news (which can be derived during the re-crawling phase). These charts show the number, portion, and the change rate of modified news for each agency. These are especially useful for finding news agencies that change and modify their news in an unacceptable manner. The time-dependent charts generally depict the influence and penetration rate of news released by each news agency in each day. These types of charts are also useful for finding the agencies which release the news in first-hand in comparison to the agencies which mostly republish the news of other agencies. Finally, in 'Reporters' section, it's possible to see the total number of news (and the number of news in each news category) published by each agency's reporter. This is especially useful for the evaluation of reporters with respect to their news quality and impact. Figure 1 shows the main segments and the complete data flow of the system.

## 4. Implementation

In order to reveal the possible barriers and to minimize the risk of the project a two phase approach based on the previous experiments has been implemented. In phase one, a pilot plan of the system was designed and implemented. To begin with, a relational database with related tables was designed. After that, as mentioned before, the major Persian news agencies were hand-picked and crawled (with focus on the recent news) and the text data was stored to the database. Other blocks of the project have also been implemented on the available data. In this phase, many shortcomings of each block were revealed and addressed accordingly. By accumulating the experiences gathered in phase 1, we encountered the problem of computational complexity. So, in the second phase, a two level database (Archive and Live) with required indexes, views and partitions were implemented in order to minimize the queries burden and maximize the speed of visualization. Other tools and methods have been also refined to enjoy the new architecture. Figure 2 illustrates a screenshot of the system.

[3] https://jqueryui.com/
[4] https://www.telerik.com/kendo-ui
[5] https://gephi.org

Figure 2: A screenshot of the system

## 5. Conclusion and future works

In this paper, we have presented the architecture, methods and approaches used to develop a robust platform for mining and monitoring the Persian news agencies. Despite the completeness of the platform and satisfaction of the end users, there is a long way ahead to implement state of the art techniques into the platform in order to be comparable to the systems in other languages.

As a work for the future, we plan to design an evaluation platform to measure the performance of the various parts of the system. We are also planning to design and implement some NLP related tasks such as hot topic detection, trend impact analysis, sentiment analysis and agency impact analysis.

## Acknowledgements

This work has been accomplished in ICT research institute, affiliated to ACECR. We want to thank Salar Mohtaj, Atefe Zafarian, Behnam Roshanfekr, Gloshan Afzali, Sepehr Arvin, and all other team members helped us in the development of the project. We also want to express our sincere gratitude to the members of Iranian Students News Agency (ISNA) for their help to the project. Special credit goes to Dr. Heshaam Faili and Mr. Khashandish for their precious guidance to the project.

## References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 363-370). Association for Computational Linguistics.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. *In advances in neural information processing systems* (pp. 856-864).

Martins, F., Magalhães, J. and Callan, J., (2016), February. Barbara made the news: mining the behavior of crowds for time-aware learning to rank. *In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 667-676). ACM.

Mohtaj, Salar, Behnam Roshanfekr, Atefeh Zafarian, Habibollah Asghari, (2018) Parsivar: A Language Processing Toolkit for Persian, *11th edition of the Language Resources and Evaluation Conference (LREC 2018)*, 7-12 May 2018, Miyazaki (Japan)

Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y. and Ngo, D.C.L., (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1), pp.306-324.

Theobald, M., Siddharth, J., & Paepcke, A. (2008, July). *Spotsigs: robust and efficient near duplicate detection in large web collections.* In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 563-570). ACM.

Volkovich, Z., Granichin, O., Redkin, O. and Bernikova, O., 2016. Modeling and visualization of media in Arabic. *Journal of Informetrics*, 10(2), pp.439-453.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4), 279-292.

Yadav, M. and Goyal, N., 2015. Comparison of Open Source Crawlers-A Review. *International Journal of Scientific and Engineering Research*, 2229, 5518, pp.1544-1551.

Zafarian, A., Rokni, A., Khadivi, S., & Ghiasifard, S. (2015, March). Semi-supervised learning for named entity recognition using weakly labeled training data. *In Artificial Intelligence and Signal Processing* (AISP), 2015 International Symposium on(pp. 129-135). IEEE.