Analysis of Coverage in Arabic Newspapers Online During the Arab Spring Through Computational Text Analysis

University Leipzig, University of Maryland

John Mathena, Tobias Wenzel, Aysha Khan, and Jacob Snyder

supervising: **Dr. Matthew Thomas Miller and Dr. Maxim Romanov**

Abstract

Our project computationally analyzes the discourse of Arabic-language news coverage published in five newspapers during the Arab Spring, comparing the frequency of topics covered based on each country's political situation during the protests and newspaper's editorial independence. We primarily achieved this using topic modeling as a computational textual analysis technique, rather than hand-coding and qualitative rhetoric-based methods used by previous studies considering similar questions. The analysis illustrated the dominance of the Egyptian Revolution on Arab Spring discourse both in Egyptian and non-Egyptian papers.

1. Introduction

Our research project uses computational methods to quantitatively compare topics covered by Arabic-language newspapers during the protests of the Arab Spring, a question that sits at the intersection of our research team members' individual experience in Arabic, Computer science, and news media studies. We chose five leading newspapers in the Middle East and North Africa and scraped the approximately 86,800 news articles they published during a two-year timeframe around the protests. We performed two rounds of topic modeling on our newspaper corpora to algorithmically determine the categories of topics each outlet covered—the anti-government demonstrations themselves, Israel and Palestine, unemployment and domestic economic issues, sports, etc.—and how much coverage was devoted to each issue at each outlet. By reading the articles most representative of each topic, we then developed a rough idea of the narrative lens each outlet used to frame these issues.

We paired these results with existing research on each newspaper's political alignment to confirm our conclusions.

This project was an exploratory analysis that lends itself to further study of the issue, through more targeted topic modeling and in-depth readings of the Arabic news coverage: the data can be compared over time; aligned with a timeline of the Arab Spring; positioned beside deeper studies of each newspaper's political and financial independence; further contextualized in existing discourse critical analysis of propaganda and rhetoric from protesters, media, civilians and political actors; or contrasted with topic modeling of either social media discourse or Western news coverage of the same protests.

2. Literature Review

2.1 Year of revolution: An overview of the Arab Spring demonstrations

The term "Arab Spring" refers to the anti-government uprisings and armed rebellions that spread throughout the Middle East and North Africa from late 2010 to 2011. In its broadest terms, the protesters advocated the overthrow of the region's various aging dictatorships and authoritarian regimes, and decried rising unemployment, economic exclusion, alienation of youth, human rights abuses, living standards, state corruption, and police violence. The first wave began in December 2010, when a young fruit vendor set himself on fire in protest before a government building in a rural Tunisian town. His self-immolation, caught on cellphone cameras and spread via social media, resonated throughout the nation as an act of civil unrest; that same day, several of the demonstrations that eventually became in Tunisia's so-called Jasmine Revolution began, and culminated in the successful removal of former President Zine El Abidine Ben Ali in late January 2011. Days after Ben Ali was ousted, mass unrest broke out in Egypt, where protesters centered at Cairo's Tahrir Square called for the removal of thirty-year President Hosni Mubarak. On February 11, 2011, Mubarak ceded power to the military. Bahrain, Yemen, Libya and Syria also saw anti-government movements begin between late January to March 2011, each with varying levels of success and facing violent state and military suppression.

With this socio-political context in mind, we analyzed five different Arabic-language, online news sources over the course of the Arab Spring to identify different trends, topics and discursive narratives presented in their coverage. For our quantitative analysis, which is primarily topic modeling in line with methods of critical discourse analysis, we collected news articles published on the websites of the following news sources between December 2010 and May 2011: Hespress, an independent Moroccan newspaper; al-Watan, an independent news source from Qatar; al-Masri al-Yawm, an independent Egyptian newspaper; al-Thawra, a daily paper owned by Syria's Arab Socialist Baʿath Party; and al-Ahram, Egypt's most widely circulated daily newspaper. Based on existing research, we also attempted to align the topics our topic modeling analysis uncover with each country's political events and each news source's editorial direction.

The Egyptian newspapers we chose exemplify that editorial dichotomy. Egypt's government owns a majority stock in al-Ahram, the country's most widely circulated daily newspaper. As the president himself appoints the paper's editor, such state-linked papers see little editorial censorship but tend to avoid publishing criticism of the state because of this affiliation.[1] On the other end is the independent and reformist Al-Masri al-Yawm (also spelled Al-Masry al-Youm), Egypt's most widely circulated privately-owned paper. Al-Masri al-Yawm is a frequent critic of the government and celebrates other challengers of the state. One analysis of both papers by Hend Selim looked at the frequency of certain words used to cover major events in the Egyptian Revolution, and found these affiliations came through in the papers' coverage and language. Until the January demonstrations, for example, al-Ahram never used the terms "demonstrations," "intifada," "revolution," "protest," "freedom" or "unemployment" before January 25; al-Masri al-Yawm used them 247 times.[2] And al-Ahram described victims as "martyrs" 25 times, versus 145 times in al-Masri al-Yawm.[3]

According to Selim, al-Masri al-Yawm's coverage aimed to be unbiased but was sometimes affected by the ousted government pressure. Her report finds al-Ahram's coverage generally tilted in favor of the Egyptian administration but took an abrupt about-face after Mubarak was removed and its circulation fell:

[1] Anti-Defamation League Arab Media Review, January-June 2012, pp. 48. http://www.adl.org/assets/pdf/anti-semitism/Arab-Media-Review-January-June-2012.pdf

[2] Hend Selim, "The Coverage of Egypt's Revolution in the Egyptian, American and Israeli Newspapers," pp. 31, Reuters Institute for the Study of Journalism

[3] Selim, pp. 41

> “It was very keen to marginalize the opposition and tarnish the demonstrators’ image. AlAhram’s coverage changed gradually to side with the people after its journalists protested against the editor-in-chief who had been appointed by the ousted president....Al-Ahram’s coverage changed completely after Mubarak’s ousting on 11th February, to side with the revolution. The military council which replaced the ousted president appointed the leaders in the stateowned [sic] newspapers and influenced their editorial policies.”[4]

Syria’s daily al-Thawra is owned by Syria’s Arab Socialist Baʿath Party and is the official newspaper of the Syrian government, alongside a few other papers controlled by the state. According to the state decree that launched the aper, its aim is to push progressive Arab society by “apply[ing] the nationalist method of socialist construction.”[5] The independent Moroccan newspaper Hespress is the country’s most popular online news and information website and the first to ever caricature King Mohammed VI. While Hespress has no official editorial line, as a collection of bloggers and journalists rather than a press institution, it has landed on the wrong side of the state: in 2008, contributor Mohamed Erraji was charged with failure to show "due respect to the king” and sentenced to two years in prison for writing an opinion piece that criticized the king for rewarding those who praise him. Independent Qatari daily Al-Watan’s chairman is royal family member Hamad bin Sahim Al Thani, and foreign minister Hamad Bin Jasim Bin Jabir Al Thani owns half of the newspaper. This is not unusual; virtually all Qatari papers are owned by branches of the royal family.[6]

2.2 LDA topic modeling and quantitative media discourse analysis

As previously noted, our ultimate data set includes eighteen months worth of articles published by five separate online news publishers, several of whom are daily outlets. Therefore, rather than using traditional qualitative methods to analyze discourses in a collection of documents or hand-coding each document for topics included, our analysis will use topic models for automated analysis of large data sets through the statistical software R.

[4] Selim, pp. 142-3
[5] Salam Kawakibi, “The Private Media in Syria,” 2010, pp. 7
[6] ADL Arab Media Review, pp. 50

Topic models are “algorithms that identify latent patterns of word occurrence using the distribution of words in a collection of documents.”[7] These algorithms produce topics, or sets of words that occur together in certain patterns, for the document set; in other words, the algorithm produces a content map of the documents by tagging words frequently used in connection with one another. Based on existing knowledge and contextual suggestions, the researcher can determine the topic that unifies each word set and interpret any internal consistencies found within the word set. Topic sets produced may include thematic issues, such as unemployment rates or government corruption, for instance. Otherwise, they may include genre and writing style, such as creative feature storied versus hard news briefs; narrative frameworks, such as framing protesters as thugs and terrorists versus activists and heroes; or events, like a specific protest or even a country’s anti-government movement broadly.[8]

Journalism researchers have found particular success with topic modeling to automatically determine patterns of news coverage from textual data, especially using the popular Latent Dirichlet Allocation (LDA) model which allows each document to be assigned multiple topics. In Jacobi, van Atteveldt and Welbers’ 2015 LDA analysis of how topics in New York Times articles dealing with nuclear technology changed over time, the topic model analysis found more specific areas of coverage rather than broader interpretive lenses. It did not, however, deduct specific viewpoints inherent in any topics as previous hand-coded analyses from 1989 did.[9] Manual interpretation, taking media, historical context, as well as the documents themselves, therefore became more important.

As part of this manual interpretation of topic model output, we will borrow methods from Critical Discourse Analysis (CDA) to help explain the construction of narratives our LDA model uncovers in the documents, placing our work within the growing field of Corpus-Assisted Discourse Studies. CDA usually concentrates on news reporting, political interviews, and counselling texts; and it goes beyond pure description of discourse analysis to interpret structural realizations and explore how and why they were produced, in a sort of

[7] Carina Jacobi, Wouter van Atteveldt and Kasper Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," pp. 2
[8] Jacobi, van Atteveldt & Welbers, pp. 2-3
[9] Jacobi, van Atteveldt & Welbers, pp. 11

"dialectical relationship with the discourse."[10] One study married topic modeling and CDA to analyze discursive depiction of the words "Muslim" and "Islam" in a 105 million word corpus comprising thirteen years of Swedish Internet forum posts. Through interpretive analysis, researchers were able to distinguish "a general discursive shift from a focus on immigrants in general, to Muslims in particular" and representations of Muslims as a conflict-ridden group through the most connected terms: terrorism, sexual abuse, violence, oppression of women.[11] The same researchers also used CDA and topic modeling to analyzing discursive connections between Islamophobia and anti-feminism in over 50 million online forum posts.[12]

2.3 Discursive studies of the Arab Spring

Much previous discursive analysis has been done in the particular context of the Arab Spring, as well—primarily looking into social media discourse, but also exploring political rhetoric and newspaper coverage—which hold significant implications for historical understandings of the Arab revolutionary moment. As Fatima Zahrae Chrifi Alaoui writes in her 2014 qualitative analysis of the vernacular discourse of Yemen's Karama Revolution:

> "this research locates a method of examining and theorizing the dialectic between agency, citizenry, and subjectivity through the study of how power structure is recreated and challenged through the use of the vernacular in revolutionary movements, as well as how marginalized groups construct their own subjectivities through the use of vernacular discourse. Therefore, highlighting the political prominence of evaluating the Arab Spring as a vernacular discourse is important in creating new ways of understanding communication in postcolonial/neocolonial settings."[13]

Tara Rhodes' 2013 qualitative discourse analysis of news outlets, blogs and social media discussions during the Egyptian and Tunisian uprisings reported fragmentation among

[10] Peter Teo, "Racism in the news: a Critical Discourse Analysis of news reporting in two Australian newspapers," pp. 12

[11] Anton Törnberg and Petter Törnberg, "Muslims in social media discourse: Combining topic modeling and critical discourse analysis"

[12] Anton Törnberg and Petter Törnberg, "Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum"

[13] Fatima Zahrae Chrifi Alaoui, "The Vernacular Discourse of the 'Arab Spring': An Analysis of the Visual, the Embodied, and the Textual Rhetorics of the Karama Revolution," 2014, pp. iii

protesters themselves, whose vocalized demand for a change in power was broad enough to encompass, in Egypt, Mubarak's leaving, the ruling party's leaving, or the entire Egyptian government's structural disintegration.[14] In Tunisia, her analysis of protesters' chants and rhetoric found a religious undertone through allusion to the shahada and to martyrdom, which unified the protesters as Shia Muslims defiant against the government. The analysis found "a dichotomous, tumultuous, and religious world, captured in the chants of both citizen groups," [15] but was limited as it was conducted in English and, as a qualitative study of hand-picked texts and slogans, was limited in scope as well. Neither limitation applies to our research, which analyzes about 86,800 articles from five leading news sources rather than focusing on topic-specific articles.

A separate 2013 analysis of news discourses in Turkish national newspapers, both pro- and anti-ruling party, looks into the narrative construction of the protests through "news actors and their quotation patterns, lexicalization, overlexicalization and syntactic preferences." This study includes a focus on both topics covered in news reports and the stories' headlines, as journalists deliberately construct headlines to "upgrade" or "downgrade" topics covered and use quotations to offer parties a chance to present their voice; however, our own research is more interested in microstructures within the text rather than macrostructures like headlines and quotes, to use their study's terminology. In their analysis on the content of these quotations, they found they were dominated by "positive commentary on the protests" and "assertions about the demise of authoritarian regimes in the Arab countries."[16] They also uncovered the lexical preferences newspapers used to present the protests as "historically, significant, widespread and widely supported," including the following:

> "massive protest; Egypt is uprising; revolution; the rage of the protesters; the waves of rebel; the collapse in Egypt; Nile revolution; secular/religious/wealthy/poor; domino effect; Tahrir victory; the fire of the rebellion; the joy of victory in Tahrir; public upheaval; public revolution; rage the people has won."

And opposition to the government was presented in such terms as the following:

[14] Tara Rhodes, "Protests in a New Perspective: A Discourse Analysis on the Arab Spring" (2013), pp 12

[15] Rhodes, pp. 18

[16] Banu Dagtas, "Constructing the "Arab Spring": News Discourses in Turkish Newspapers"

> “go away—go away; the last trump cards of Mubarak, the streets toppled Mubarak; the end of the bobo doll of the Middle East; the strongholds of Gaddafi have fallen down; Gamal, take your father with you and go away; enough is enough Mubarak; God damn Mubarak!; the end of the road; the last pharaoh; Benghazi butcher; the bluff for chaos; 30-year reign; insanity of power; the last shock to Mubarak is by Obama; the death squads of Gaddafi; the last flutters of Gaddafi (Sabah), Mubarak, go away!; the Mubarak era is ending; Gaddafi splits up European Union (Zaman).”[17]

A similar, hand-coded qualitative discursive study of how the Egyptian uprising was framed in Arabic newspapers (including both semi-official papers and independent papers) and social media, found that state-run newspapers framed the uprisings as “a conspiracy on the Egyptian state, warning of economic consequence and attributing blame and responsibility for the chaos on others.”[18]

While much qualitative research has been done into narrative discourses produced in specific countries and movements during the Arab Spring, and other computational research has used corpora of Twitter and blog posts to analyze social media discourses during the same time period (the R-Shief archive, for example, holds years of Twitter and Facebook data surrounding the 2011 uprisings)—our study is unique in it that it offers a preliminary *computational* analysis of online news media across countries. Ours is also distinct in that it focuses particularly on topics covered, rather than focusing on the rhetoric; our analysis also includes all news stories, not just Arab Spring-related articles, from news outlets from several countries. Much of the existing research available in English has also been limited to English-language news coverage and social media.

3. Documentation and Methodology

The following section describes the steps of scraping (4.1) and cleaning (4.2) as well as the exploratory data analysis (4.3, 4.4). In this context scraping is the step of automated downloading raw HTML files from a specific news source’s website. The requests are followed by a pause to limit prevent the target server from overloading. Cleaning describes the extraction and reformatting of selected text elements, i. e. the article itself, its

[17] Dagtas, pp. 27

[18] Naily Hamdy and Ehab Gomaa, "Framing the Egyptian Uprising in Arabic Language Newspapers and Social Media Authors"

title or date it was published. Although it is possible to combine the steps and directly extract the article, we chose this method to be able to access the raw HTML files in case of errors during the scraping process.

Since this project aims to do research not on one specific newspaper website's homepage but rather a whole selection (2.1), the process of scraping homepages and extracting data or cleaning differs for each site. However, as the method largely stays the same, it thus can be explained using pseudo-code. The commented source code can be on our Github project page[19]. The code was mainly written in R inside the integrated development environment RStudio[20].

3.1 Scraping

Algorithm 1: *scrape.day.$NEWSPAPER*. The algorithm navigates to the main page of the news-source, saves all relative and direct links and saves the articles as HTML in a target folder.

Input:

Date or direct link

Output:

Raw HTML files

```
if date as input:
    main ← load main page
    article.links ← extract links(main)
    for link in article.links do
        save as HTML
    end for
else if direct link as input:
    article ← load article
    save(article.html)
end if
```

[19] Available in full at https://github.com/Islamicate-DH/hw/tree/master/newspaper_group/code

[20] R version: 3.2.3 (2015-12-10). RStudio: Version 1.0.44 – © 2009-2016 RStudio, Inc. Linux x86_64

The R packages used during scraping and cleaning are mainly rvest[21] and tidyr[22]. Rvest provides functionalities to download and examine homepages as XML. Tidyr is used to pipe the output of one R command into another command, usually to narrow down XML elements in a given homepage. This makes the code shorter and easier to read.

After loading the packages and the functions implemented in D.1, a time sequence is set which is to be downloaded from the specific news source. This sequence consists out of date strings. In the following, the function scrape. $NEWSPAPER (Algorithm 1) is called with sapply. sapply applies a function on each element of a vector, in this case holding links to homepages. Depending on the homepage structure we first navigate to the main page of the newspaper and collect all links to articles written on this day. In the next step these links are loaded as XML and the resulting homepage is saved in a target folder. Errors occurring during the scraping are noted in a log file.

3.2 Text Extraction

Algorithm 2: *clean.$NEWSPAPER*. The Algorithm takes a source folder as input and extracts text elements which hold the article and date. The result is appended to a csv file.

Input:
source folder

Output:
csv file

```
for homepage in source folder do
    article.homepage ← load html
    dateString ← getDateElements('html-node')
    year, month, day ← extractDate(dateString)
    article ← getArticleElements('html-node')
    save as csv(year, month, day, article)
end for
```

[21] T. Harvest, W. Pages, D. Wrappers et al. “Easily Harvest (Scrape)Web Pages”

[22] Hardley Wickham. “Easily Tidy Data with `spread()` and `gather()` Functions”

To keep the data format as easy as possible, we only extracted each article and corresponding date. This enables us to identify each article with a uniform resource identifier (URI) consisting out of a short form of the newspaper as shown in Table 1, the date as concatenated string without separators and an index starting at the first day of our chosen timespan. HP20101201$1 identifies an article written December 1, 2010. Al-Masri Al-Yawm only uses Arabic numerals making it difficult to extract them as numbers. In this case only an index is used to identify an article. Other pages like Ahram used month names instead of numbers, thus had to be converted using a self-written function f.replaceMonthNames. The identifiers and the articles are saved in the cross platform format comma-separated value (csv).

TABLE 1: News source corpus, with abbreviations.

Newspaper	**Short form**	**Corpus** *[articles]*	**AVG article size** *[characters]*	**AVG article size** *[words]*
Hespress	HP	54368	268	44
al-Watan	AL	1749	1962	328
Ahram	AH	8708	1622	273
al-Masri al-Yawm	AY	19976	2319	390
Thawra	TH	2016	2110	318
Total	-	86817	-	-

Having saved all corresponding data in the target folder, the data is ready to be extracted. The extraction function expects a source folder. Each file in this folder is first loaded as XML into the R environment. Then the script navigates to preselected HTML nodes and extract both the article fragments and the date. If the page was indeed an article[23],

[23] Not all contain articles. Some of them are links to do not contain interpretable information. RSS feeds (Rich Site Summary) are not saved, either.

the elements are appended to a target csv file. Any files which do not contain an article are noted in a log file.

The final renaming, which results in the format explained previously, is now applied. This was done in an separate step, because we decided to narrow down the data timespan and adjust the URI format.

3.3 Word frequency analysis

We utilized two visualizations in addition to a script using the ‘tm‘ package[24] in R as our main methods of Word frequency analysis and comparison. Those two visualizations are histograms of a selection of the 100 MFWs (most frequent words) as well as word clouds of the 100 MFWs, which we put in the appendix for reference.

We chose the words for the histogram by manually eliminating the top 40 words in each frequency list of each paper. We did this for two reasons. Firstly, if we simply plotted the MFWs including those top 40 each paper would look rather similar and that would not aid our analysis. Secondly, our efforts to automatically extract stop words (such as حيث ,ال ,أن, etc.) proved unfruitful. The usefulness of the histograms based on the most frequent words in each newspaper was limited. However, this data was useful when compared to the topic modeling data that we analyzed. The most frequent word data helps to reaffirm which topics were discussed the most in each newspaper and gives a basic insight into some of the points of focus for each newspaper. Although the most frequents word analysis did not lead to further analysis, it played a relevant role when we examined the topic modeling data.

3.4 Topic modeling

The procedures used during topic modeling are borrowed from a script to study Persian text corpora by Thomas Köntges (University of Leipzig). The packages mostly used are lda[25] for topic modeling and LDAvis[26] for its visualization.[27] We used a limited amount of

[24] Ingo Feinerer, Kurt Hornik and Artifex Software, Inc. Text Mining Package. Version 0.6-2

[25] Jonathan Chang. Collapsed Gibbs Sampling Methods for Topic Models. Version 1.4.2

[26] Carson Sievert and Kenneth Shirley (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on interactive Language Learning, Visualization, and Interfaces, p.63-70

[27] The script uses a Web API to stem Persian words. Stemming describes the process of the finding roots of words to reduce the text corpus and remove words which have the same or very similar

stopwords to decrease the usage of prepositions and articles. After reading in a text corpus holding one news source punctuation, numbers and remaining control characters are removed from the text. Finally the texts are split up into words and their frequency is calculated and the put into the required format required by lda. To get a specific model one can choose in how many topics the algorithm should divide the corpus vocabulary and how many iterations it is supposed to run. We decided to use 15 topics each holding 40 words characterizing the topic. The fitting is iterated 5000 times.

After the fitting process is done the data is parsed to json in order to display it correctly within the visualization plugin LDAvis (Figure 15: Example output of LDAvis). The results can be examined in a browser.[28] The interface allows users to walk through the different topics shown as spheres. The right side indicates the number of times the terms appear in the corpus or a specific topic. Alternatively the words can be selected to show the topics they appear in. This allowed us to identify and compare subjects in the way they were talked about in the given time period.

To qualify as well as to have numbers to substantiate our findings we looked at the JSON file created at the end of the topic modeling process. It holds the relative values for each topic, i.e. the probability of the topic given the corpus (Table 2). Additionally we found representative articles by maximizing the probability of one topic. This probability of each article given a topic is saved in a csv file called theta. Its columns hold the topic consisting out of words and the rows are the URIs, created during the data preparation steps. Then maximizing can then be achieved by ordering the table in relation to one topic column in an office spreadsheet program.[29] The article which holds the highest value was then chosen to be representative.

TABLE 2: Probability of each topic. Short forms like in TABLE 1. The name of the topic is represented by seven words which are characteristic for the topic.

source	topic	name of topic	frequency [%]	Arab Spring topic frequency

meaning. In Arabic this procedure is more difficult than in Persian. One possible R stemming package is ArabicStemR, which allows stemming as well as transliteration.

[28] The browsers Internet Explorer and Google Chrome do not display the files. Use Firefox instead.

[29] We used Libre Office Calc Version: 5.1.4.2.

				[%]
AH	1	مصر الثورة ولكن ليس لأن مثل يجب	18.02	45.61
AH	3	المجلس الشعب مجلس اللجنة القانون الدستور البرلمان	7.95	
AH	5	الشرطة المسلحة الثورة القوات المتظاهرين الأمن يناير	7.67	
AH	6	مصر المصرية الخارجية المصري العربية الأمريكية الدول	7.37	
AH	11	الانتخابات حزب الإخوان اللجنة المرشحين الانتخابية الحزب	4.60	
AY	1	مصر الثورة النظام أى هى الدولة الآن	16.52	39.30
AY	3	المجلس مجلس وزير وزارة أمس الوزراء المصرى	7.38	
AY	7	الشرطة الأمن الثورة التحرير المتظاهرين ميدان مبارك	6.51	
AY	8	الحزب الانتخابات الشعب الوطنى مجلس الإخوان المجلس	5.89	
AY	15	القذافى أمس ليبيا تونس الرئيس وفى البلاد	3.00	
AL	9	الأميركية العراق القوات الأميركي الجيش المتحدة العراقي	5.56	
AL	11	لبنان المحكمة اللبناني الدولية اللبنانية الرئيس مجلس	4.59	
AL	12	السودان الرئيس الانتخابات العاج المصري غباغبو ساحل	3.68	
AL	13	الأميركية الحرب طالبان باردو الاستراتيجية أفغانستان العسكرية	3.55	
HP	3	السياسية السياسي المغرب الحركة الحزب حزب حركة	11.81	
HP	5	الحكومة حزب بنكيران والتنمية العدالة عبد الحزب	5.93	
HP	7	مدينة البيضاء الشرطة أمس السلطات المحكمة أحد	5.70	
HP	13	المغرب المغربية الصحراء البوليساريو الجزائر الملك المتحدة	4.71	
HP	14	المغربية الإعلام الصحافة موقع جريدة المغرب القناة	4.08	
TH	2	سورية الوطنية الرئيس الشعب السوري الأسد السورية	9.40	
TH	4	ليبيا الليبية الليبي أمس وكالة الشعب مصر	8.74	
TH	7	حالة قانون أمن الدولة الطوارئ القانون التظاهر	6.07	
TH	8	لبنان اللبنانية اللبناني الله السورية المقاومة السوريين	5.43	
TH	12	المواطنين الامن النار المسلحة عناصر وأضاف مدينة	4.26	
TH	14	بغداد بجروح الشرطة العراقية مقتل مصدر العراق	3.97	

4. Visualization and Analysis

We performed two topic modeling analyses on our newspaper corpus. The first run analyzed each newspaper corpus individually giving us 15 different topics, or 75 topics in total, for each newspaper allowing us to compare what each country covered during our time period. The second run treated all the articles from all five newspapers as one corpus. Through this method, we determined the topics that appeared in all the papers and to what extent each country focused on each of those topics. With both forms of topic modeling we were able to analyze our data in several different ways which allowed us to more thoroughly examine the coverage of each news source throughout the duration of the Arab Spring.

4.1 Results of Topic Model 1: Analysis of topics within countries

1) Egyptian papers
    a) One topic on the constitutional reform committee
        i) Arabic keywords from top 100 MFWs: الدستور, اللجنة, المجلس
    b) One topic on the clashes between protesters and the police
        i) Arabic keywords from top 100 MFWs: الشرطة, المتظاهرين
    c) One opinion topic on foreign affairs during Arab Spring
        i) al-Ahram – post-Arab Spring election topic
        ii) al-Masri al-Yawm – Models of social change topic
2) Al-Watan (Qatar)
    a) Some topics on domestic affairs unrelated to the Arab Spring
        i) Arabic keywords from top 100 MFWs: القطر, العام, الثقافة
    b) One topic on U.S.-Middle East affairs
        i) Arabic keywords from top 100 MFWs: الأمريكي , الاستراتيجية, العسكري
    c) One to two topics on Arab Spring protests in other MENA countries
        i) Libya
        ii) Sudan
        iii) Lebanon

iv) Egypt

4) Hespress

a) One topic focusing on constitutional reform in Morocco
    i) Arabic keywords from top 100 MFWs: السياسة، حركة، المغرب
b) One topic focusing on violations of the law and safety within Morocco
    i) Arabic keywords from top 100 MFWs: السلطان, الشرطة, الأمن
c) Many other topics regarding domestic affairs within Morocco

5) Thawra

a) One topic on Palestinian and Israeli relations and the conflict between them
    i) Arabic keywords from top 100 MFWs: فلسطين, قانون, الأمن
b) One topic about changes in the political parties in other countries in the Arab world
    i) Arabic keywords from top 100 MFWs: شعب , وزير، الخارجية
c) One topic regarding the political foundations of various countries in the Middle East
    i) Libya
    ii) Syria
    iii) Lebanon
    iv) Egypt

6) Arab Spring topic frequency (% out of 100): How much topic space was devoted to domestic topics on the Arab Spring in each paper

v) Ahram – 45.612
vi) al-Masri al-Yawm – 39.304
vii) al-Watan – 17.385
viii) Hespress – 32.237
ix) Thawra – 19.723

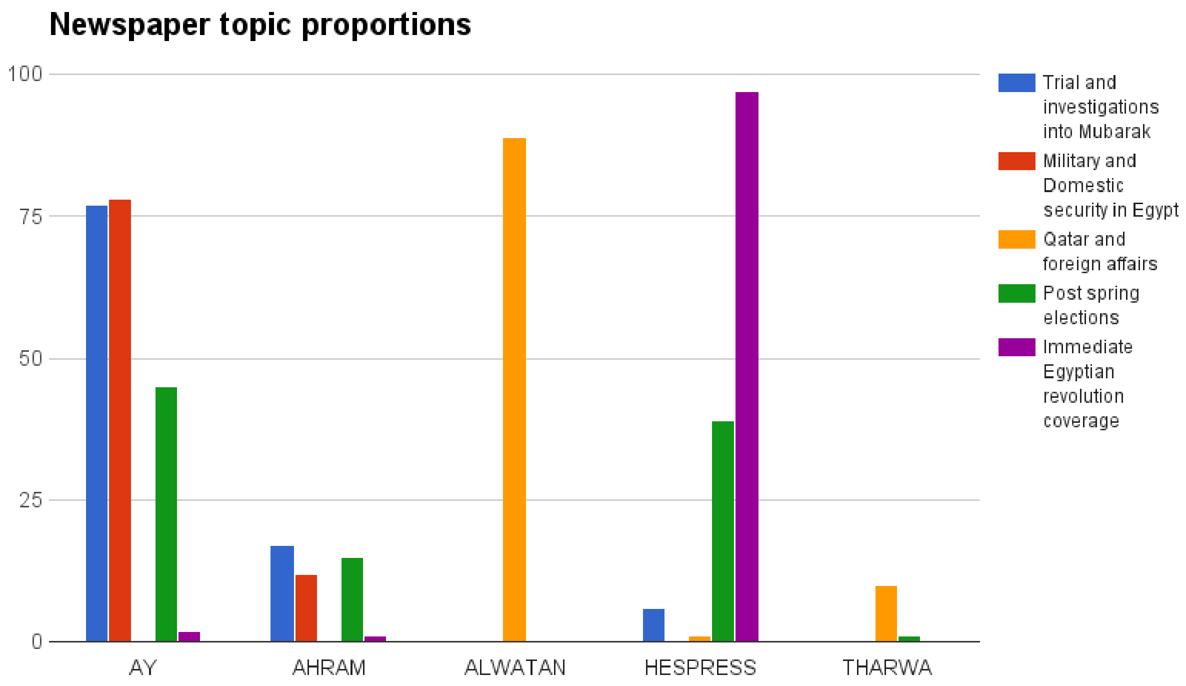
In countries with an actual revolution, so to speak, there appears to be more document space allocated to discussing the political consequences of said revolution on that country. In Egypt, that discussion revolved around the quickly formed constitutional reform committee and the violence perpetrated by the late Mubarak's regime against civilian protesters. **Al-Masri al-Yawm** and **Ahram** also dedicated space for opinion pieces discussing models and goals for the socio-political change in the country in addition to the post Arab Spring polarization and overabundance of political factions. The political congestion and division eventually resulted in a divided parliament and ousting of the first democratically elected president in Egyptian history, Mohammad Morsi.

Due to their different editorial structures and ownerships, **Ahram** featured few articles criticizing the government with only one topic containing any opinion articles, while **Al-Masri al-Yawm** contained much more detailed opinion pieces that actually came out during the Spring. Through a close reading of the most representative articles in those topics, it became clear that these articles were opinion articles. One article from AY for example, cited the works of Hannah Arendt, a political theorist, on how power affects the political process. Another article lays out what a successful democracy must do in addition to stating that the opposing forces (قوى المعارضة المختلفة) and political factions must find common ground in order to move the country forward.

Al-Watan in Qatar still dedicated topic space to the Spring but through a foreign lens. Articles from al-Watan examined the Arab Spring protests, demonstrations, and clashes happening in other Arab countries, as they themselves did not undergo a regime change. Additionally, the paper is owned by a Qatari businessman loyal to the Royal family and the Foreign minister owns half the paper, which explains their reluctance to report on too many domestic political issues. However, the fact that a significant amount of topic space continued to inform readers of the struggles fellow Arabs faced on the ground in their home countries illustrates the widespread and overarching effect the movement had on each and every country in the Middle East.

Hespress in Morocco, like the Egyptian papers, reported heavily on domestic events such as the efforts to reform the constitution and democratize the Moroccan government. As this paper is truly independently owned and run, their focus on domestic issues is not surprising. Additionally, people felt the Ministry of Justice needed to be more transparent in order to address some human rights issues in the country. Generally, most articles focused on discussing the protests through debates and fairly objective thought pieces. There was limited reporting on foreign events and not many articles appeared extreme in their reporting using loaded words or over-reporting from a single side. Intriguingly, some articles discussed how other newspapers in Morocco covered the events.

Thawra in Syria, like al-Watan, reported the events unfolding in the Arab Spring through a mostly foreign lens, largely due to the fact the ruling political party (Ba'th party) owns the paper. The paper covered how the political standing of countries changed because of the Arab Spring but a large portion of topic space reported news regarding the Israeli-Palestinian conflict. This is was not uncommon as each paper had a topic devoted to the issue but that topic alone accounted for 15 percent of front page space, the most of any paper in our corpus. The paper did discuss the bubbling tensions between Assad's regime and the public to some degree but the overall focus of Thawra was external.



4.2 Results of Topic Model 2: Analysis between countries

1) 1. Trial and investigations into Mubarak topic
   a) Arabic keywords from top 100 MFWs: المحكمة, التحقيقات, مبارك

b) Newspaper proportions (percent of top 100 articles belonging to each paper)

i) al-Masri al-Yawm – 77

ii) al-Ahram – 17

iii) Hespress - 6

iv) al-Watan – 0

v) Thawra – 0

c) Topic description

i) This topic mostly discusses the proceedings of the trial of Mubarak and his government after the revolution toppled his regime. Much of the talk described the length of sentence and the accusations (mostly finances judging by the MFWs). The articles are objective and free of editorializing and/or loaded words.

2) Military and Domestic security in Egypt topic

a) Arabic keywords from top 100 MFWs: الأمن ,الشرطة ,اللواء

b) Newspaper proportions (How many of top 100 articles came from each paper?)

i) al-Masri al-Yawm – 78

ii) al-Ahram – 12

iii) Hespress – 0

iv) al-Watan – 0

v) Thawra – 0

c) Topic description

i) This topic describes sentences and and on-going cases in Egypt ian courts including the trial involving the former Egypt ian president, Hosni Mubarak.

3) Qatar and foreign affairs topic

a) Arabic keywords from top 100 MFWs: الخارجية ,الدول ,قطر

b) Newspaper proportions

i) al-Watan – 89

ii) Thawra – 10

iii) Hespress – 1

iv) al-Masri al-Yawm – 0

v) al-Ahram – 0

c) Topic description

i) This topic discusses inclusively qatari foreign affairs many of which focus on Qatar’s hosting on the world cup in 2022.

4) Post Spring elections topic

a) Arabic keywords from top 100 MFWs: حزب, الانتخابات, المجلس, الإخوان

b) Newspaper proportions

i) al-Masri al-Yawm – 45

ii) Hespress – 39

iii) al-Ahram – 15

iv) Thawra – 1

v) al-Watan – 0

c) Topic description

i) This topic contains discourse on constitutional reform and post revolution elections.

5) Immediate Egyptian Revolution coverage

a) Arabic keywords from top 100 MFWs: مص ,الثورة, الآن

b) Newspaper proportions

i) Hespress – 97

ii) al-Masri al-Yawm – 2

iii) al-Ahram – 1

iv) Thawra – 0

v) al-Watan – 0

c) Topic description

i) From al-Masri al-Yawm – nothing currently happening will make the future any better than the past. We have seen all of this before. If corruption occurred under the current laws and the laws themselves do not change then how is the new government going to be any different. Ahram contribution is insignificant.

Al-Masri al-Yawm and **Ahram** obviously contributed the most to topics on events in Egypt such as the Mubarak trial and subsequent post revolution election. It was interesting to see how much the Egyptian papers dominated some topics not directly pertaining to Egypt ian issues like the election and constitutional reform topic. While not a uniquely Egyptian issue, it is clear that the Egyptian narrative on the Arab Spring is very influential.

As we saw from the individual newspaper analysis, **al-Watan** dedicated the least amount of coverage to actual discourse surrounding the Arab Spring that was not through a foreign lens. Taking that into consideration when analyzing this second round of results, it makes sense that the only topic to which al-Watan makes a significant contribution is the topic dedicated to Qatari affairs. Even that topic is mostly comprised of summits and meetings between Qatar and other Arab nations.

Hespress contributed a large amount of news space to report on the Egyptian Revolution as it occurred. The news source also allocated a large amount of space to discuss elections that were taking place after the Arab Spring had ended. This is unsurprising considering that the effects of the Arab Spring were much more limited in Morocco and these statistics validate that Hespress' discussion on the Arab Spring was mainly external.

Using this form of topic modeling it is clear that **Thawra** had a very limited engagement with these topics. It also shows that while there was a large amount of coverage on foreign affairs in the Arab world during the Arab Spring, the paper had a very limited discourse on the events within Egypt during the revolution.

4.3 Summary of findings and discussion

Our topic analysis within each country illustrated each country's individual coverage of the Arab Spring as well as what role ownership played in the topics covered by the papers we examined. Government owned and influenced papers (Tharwa and Al-Watan) tended to focus on foreign issues because of their inability to report on domestic issues. The lone exception appears to be Ahram, which was government owned before the start of the revolution, but flipped its support following the ousting of Mubarak in 2011.

While similar topics emerged in each country (coverage of demonstrations, post revolution elections, etc.), this initial analysis revealed each country's coverage as either internal, focusing on domestic issues resulting from the movement, or external, examining issues in other countries more often than issues domestically. The Egyptian papers (Ahram and al-Masri al-Yawm), as well as Hespress (Morocco), were internally focused while al-Watan (Qatar) and Tharwa (Syria) were more externally focused.

The second round of topic analysis allowed us to measure each newspaper's contribution to the same 15 topics. From this analysis, we saw that the Egyptian papers dominated discourse surrounding the key Arab Spring related topics, both in Egypt and across the Middle East. While each paper also dominated the topic related to its country (Hespress and the Moroccan topic, al-Watan and the Qatar topic).

Taken as a whole, these analyses demonstrate that the Egyptian Revolution and its fallout drove the discourse surrounding the Arab Spring across the Middle East. This is evident given the Egyptian newspapers focused on domestic issues to a greater extent than any other paper and that other papers dedicated significant space to discussion of Egyptian issues. Hespress, for example, accounted for 97 of the top 100 most related articles to a topic surrounding on-going coverage of the Egyptian Revolution as it took place. Additionally, articles from non-Egyptian papers use biased language such as describing those killed in protests as "martyrs," indicating support of the revolution and illustrating the strong emotions evoked in non-Egypt ian Arab countries as a result of the spring.

The explanation for this abundance of discourse and discussion surrounding the Egyptian revolution across the MENA region should be the subject of further study. Two potential causes of these findings are laid out here: Timing and exposure. The Egyptian ousting of Hosni Mubarak was the second regime toppling in less than a month in January 2011, as well as the second in along line of regime changes to occur across the region. The sight of their most populous and iconic nation undergoing such a change had a very profound, emotional effect on Arabs across the region. Additionally, Egyptian affairs are very salient across the Middle East given that even Turkish News coverage of the Arab Spring was dominated by the Egyptian revolution[30].

Future research into the causes of this findings could utilize social networking analysis in combination with Internet penetration and social media usage measures to

[30] Dagtas, pp. 27

determine how news about Egyptian affairs spreads differently across the Middle East, especially through social media outlets like Whatsapp and Facebook.

5. Conclusion

Before we discuss the big picture of this investigation we will lay out some limitations and some future directions for our research. Many of our limitations were in part caused by our project deadline and limited work time.

Firstly, we were only able to use newspapers that were easily scrapable therefore the newspapers used are not completely representative of that country's media. In Egypt, we were able to obtain both state-controlled and independent newspapers and this is the ideal method for any country used in this kind of analysis but we were limited by scraping compatibility and the time we had to work on the project.

Secondly, we had to limit our time frame because we did not have the resources to analyze all of the data we got from our original time frame of 2010 - 2012. We also had nearly two dozen potential newspapers from additional countries. We ultimately decided to restrict our time frame from December 2010 to April 2011, to make the amount of analysis needed more feasible given the deadlines of our project.

Lastly, as stated in the Discussion, this investigation illustrated the large role the Egyptian Revolution played in discourse surrounding the Arab Spring across the countries and papers we analyzed. Each paper devoted significant front page space to Egyptian coverage and the Egyptians paper articles were highly representative of topics relating to the Spring.

In future analyses, we would like to add papers from additional countries, both independently and state owned, to confirm our findings. Future topic models could be run solely on the articles relating to Arab Spring topics excluding articles from Sports or art topics. This analysis would yield more precise descriptions of discourse surrounding the movement. We would also like to investigate the cause for the dominance of Egyptian news over Arab Spring discourse.

Appendices

A. Sources

Anti-Defamation League Arab Media Review, January-June 2012, pp. 48. URL http://www.adl.org/assets/pdf/anti-semitism/Arab-Media-Review-January-June-2012.pdf

Hend Selim, "The Coverage of Egypt's Revolution in the Egyptian, American and Israeli Newspapers," pp. 31, Reuters Institute for the Study of Journalism

Salam Kawakibi, "The Private Media in Syria," 2010

Carina Jacobi, Wouter van Atteveldt and Kasper Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling"

Peter Teo, "Racism in the news: a Critical Discourse Analysis of news reporting in two Australian newspapers," pp. 12

Anton Törnberg and Petter Törnberg, "Muslims in social media discourse: Combining topic modeling and critical discourse analysis"

Anton Törnberg and Petter Törnberg, "Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum"

Fatima Zahrae Chrifi Alaoui, "The Vernacular Discourse of the 'Arab Spring': An Analysis of the Visual, the Embodied, and the Textual Rhetorics of the Karama Revolution," 2014, pp. iii

Tara Rhodes, "Protests in a New Perspective: A Discourse Analysis on the Arab Spring" (2013), pp 12

Banu Dagtas, “Constructing the “Arab Spring”: News Discourses in Turkish Newspapers”

Naily Hamdy and Ehab Gomaa, "Framing the Egyptian Uprising in Arabic Language Newspapers and Social Media Authors"

T. Harvest, W. Pages, D. Wrappers et al. Easily Harvest (Scrape)Web Pages Description. URL https://cran.r-project.org/web/packages/rvest/rvest.pdf, Version 0.32. , 2016-06-17

Hardley Wickham. Easily Tidy Data with `spread()` and `gather()` Functions. URL https://cran.r-project.org/web/packages/tidyr/tidyr.pdf , Version 0.6.0 ,2016-08-12

Ingo Feinerer, Kurt Hornik and Artifex Software, Inc. Text Mining Package. URL https://cran.r-project.org/web/packages/tm/tm.pdf, Version 0.6-2

Ian Fellows. Word Clouds. URL https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf, Version 2.5

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Hadley Wickham (2016). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0. URL https://CRAN.R-project.org/package=stringr

Jeroen Ooms (2016). curl: A Modern and Flexible Web Client for R. R package version 2.3. URL https://CRAN.R-project.org/package=curl

Steve Lianoglou, Jim Nikelski, Kirill Müller, Peter Humburg and Rich FitzJohn. (2015). optparse: Command Line Option Parser. R package version 1.3.2. URL https://CRAN.R-project.org/package=optparse

Duncan Temple Lang and the CRAN Team (2016). XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98-1.5. URL https://CRAN.R-project.org/package=XML

Duncan Temple Lang and the CRAN team (2016). RCurl: General Network (HTTP/FTP/...) Client Interface for R. R package version 1.95-4.8. URL https://CRAN.R-project.org/package=RCurl

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Carson Sievert and Kenneth Shirley (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on interactive Language Learning, Visualization, and Interfaces, p.63-70

Jonathan Chang. Collapsed Gibbs Sampling Methods for Topic Models. URL https://cran.r-project.org/web/packages/lda/lda.pdf, Version 1.4.2, 2015-11-22

B. Visualizations and Tables

B.1 Most frequent word clouds

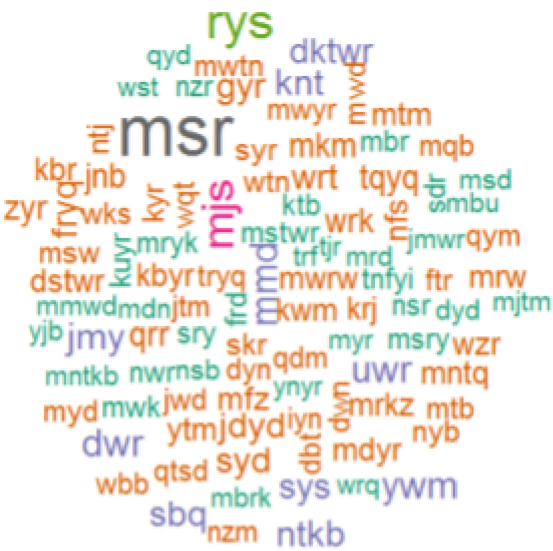


Figure 1. Transliterated word cloud of Ahram MFWs.

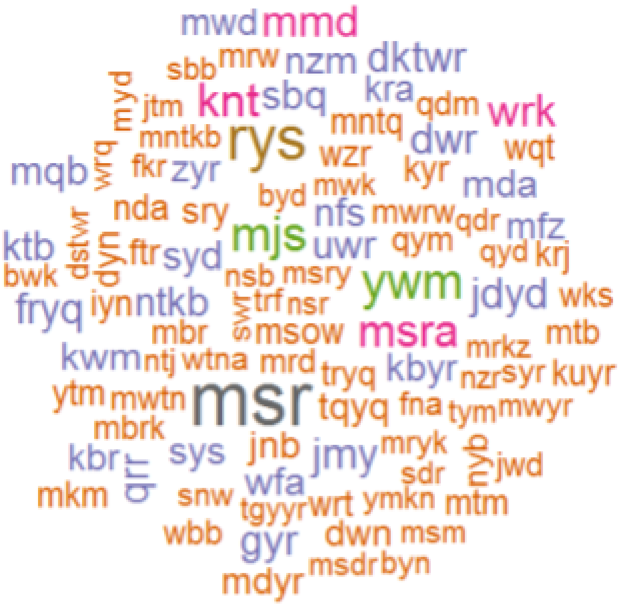


Figure 2. Transliterated word cloud of
Al-Masri Al-Yawml MFWs.

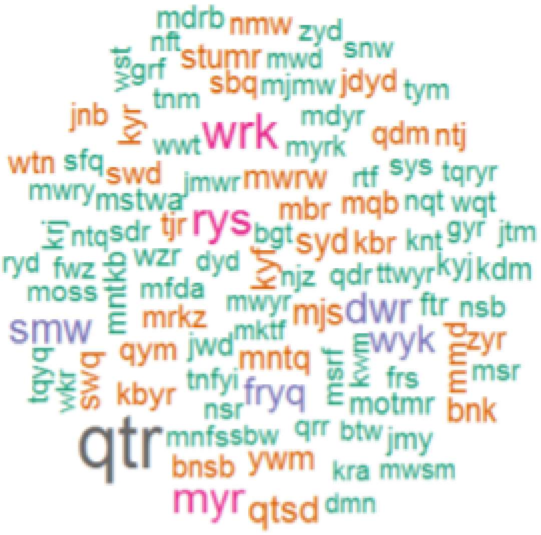


Figure 3. Transliterated word cloud of
Al Watan MFWs.

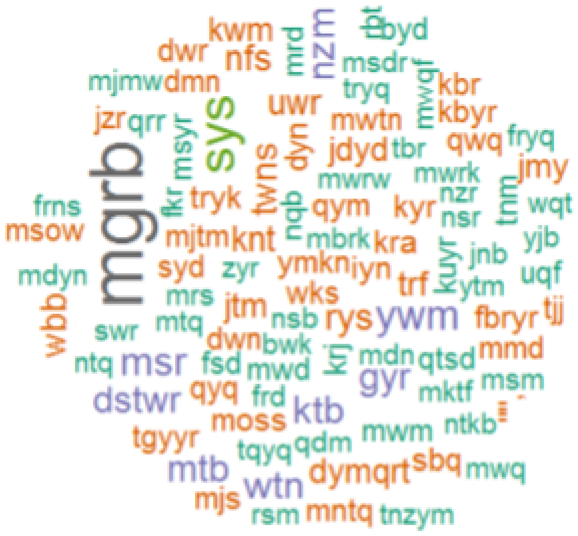


Figure 4. Transliterated word cloud of
Hespress MFWs.

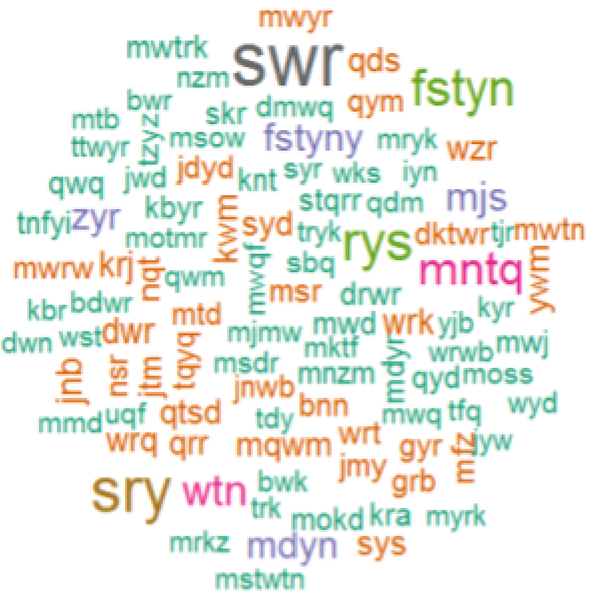


Figure 5. Transliterated word cloud of Thawra MFWs.

B.2 Histograms

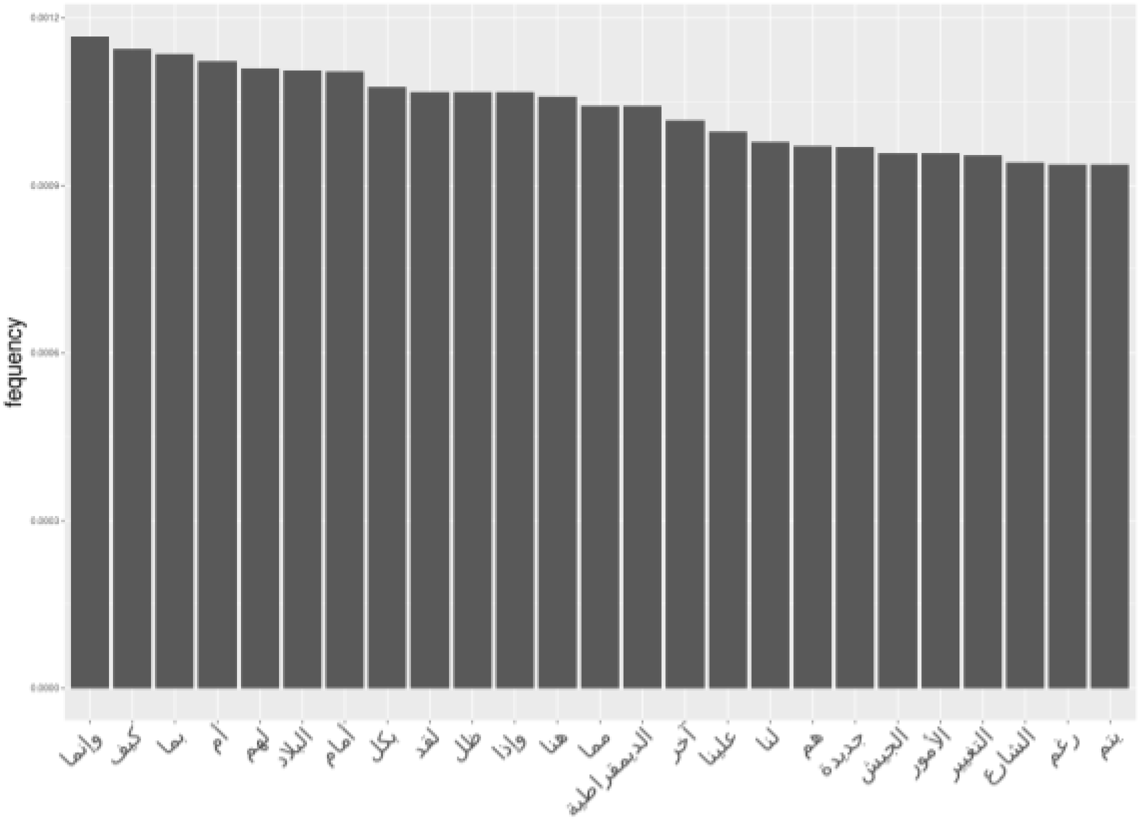


Figure 6. Most Frequent Words by Newspaper.
Egyptian revolution coverage.

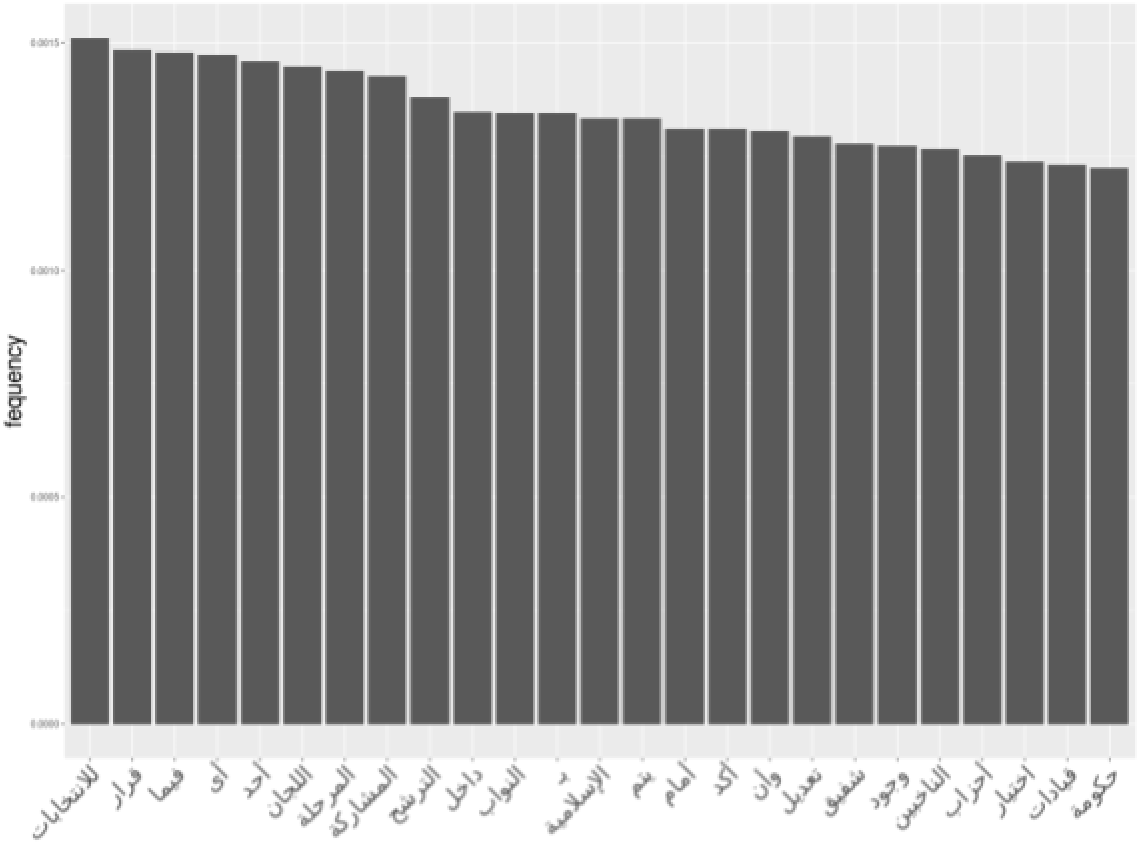


Figure 7. Most Frequent Words by Newspaper.
Post Arab Spring Elections.

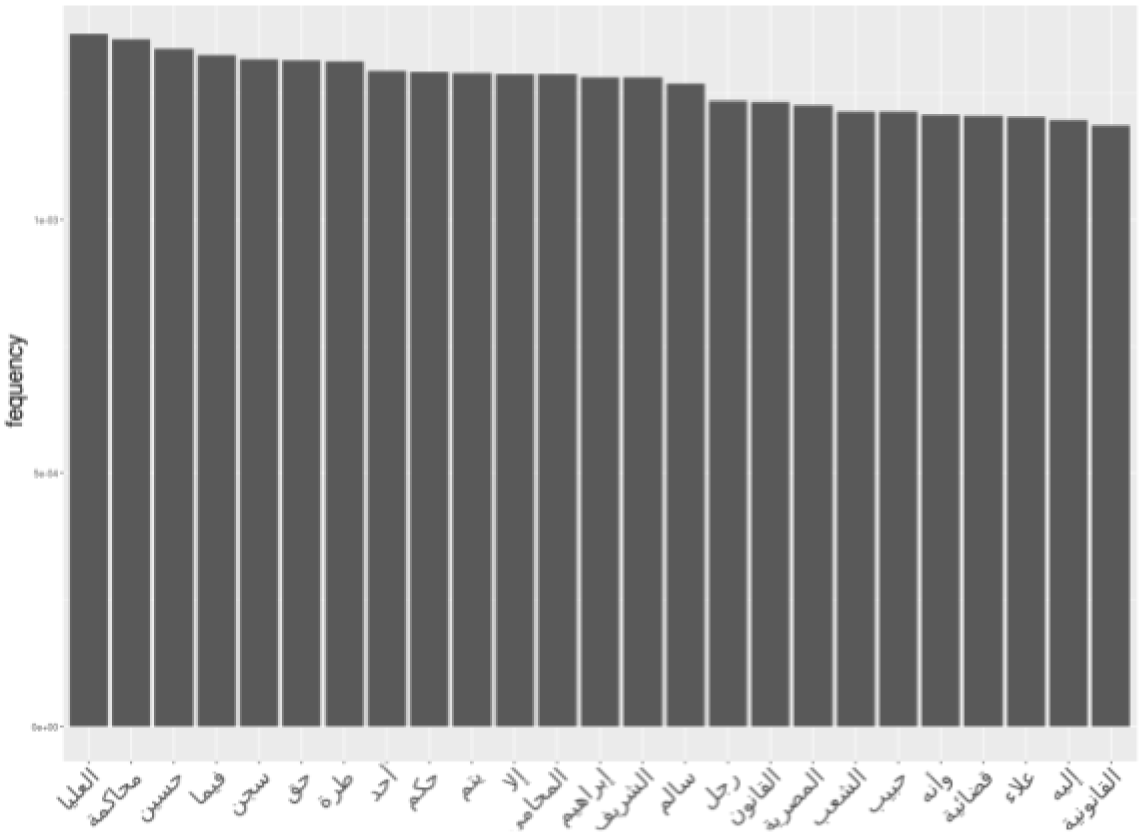


Figure 8. Most Frequent Words by Newspaper.
Mubarak trial coverage.

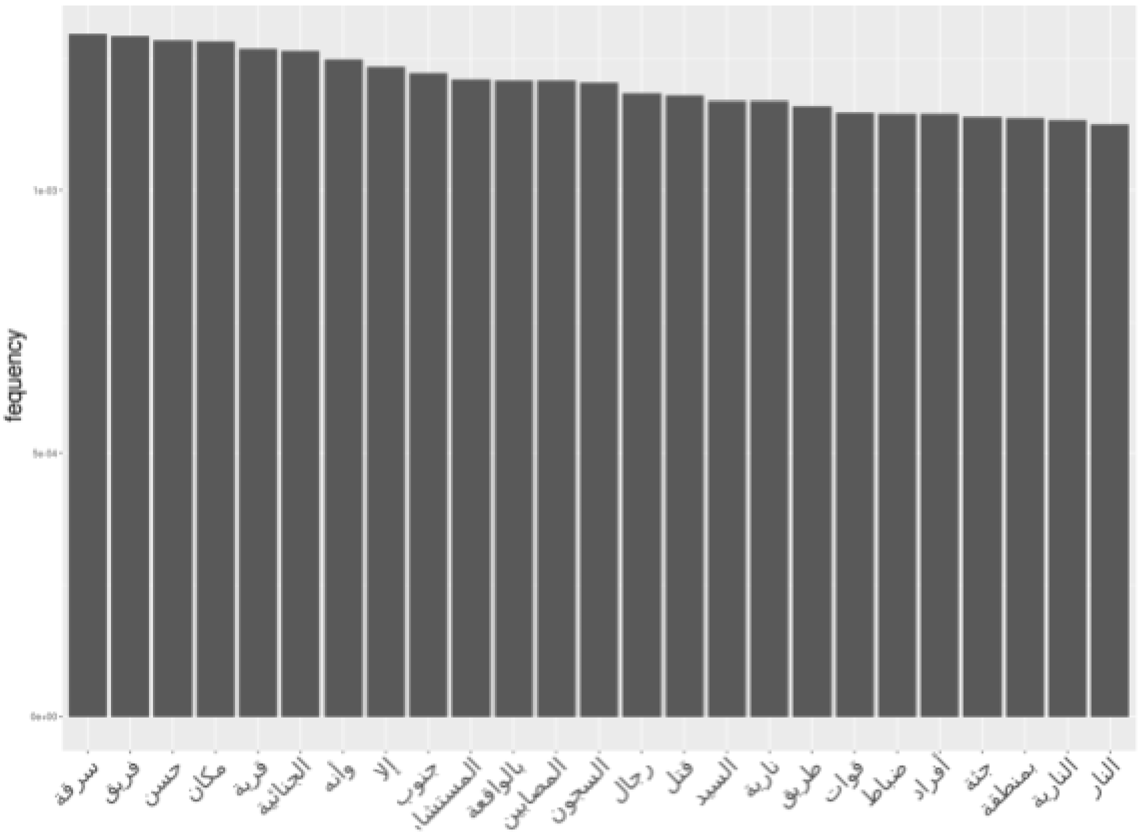
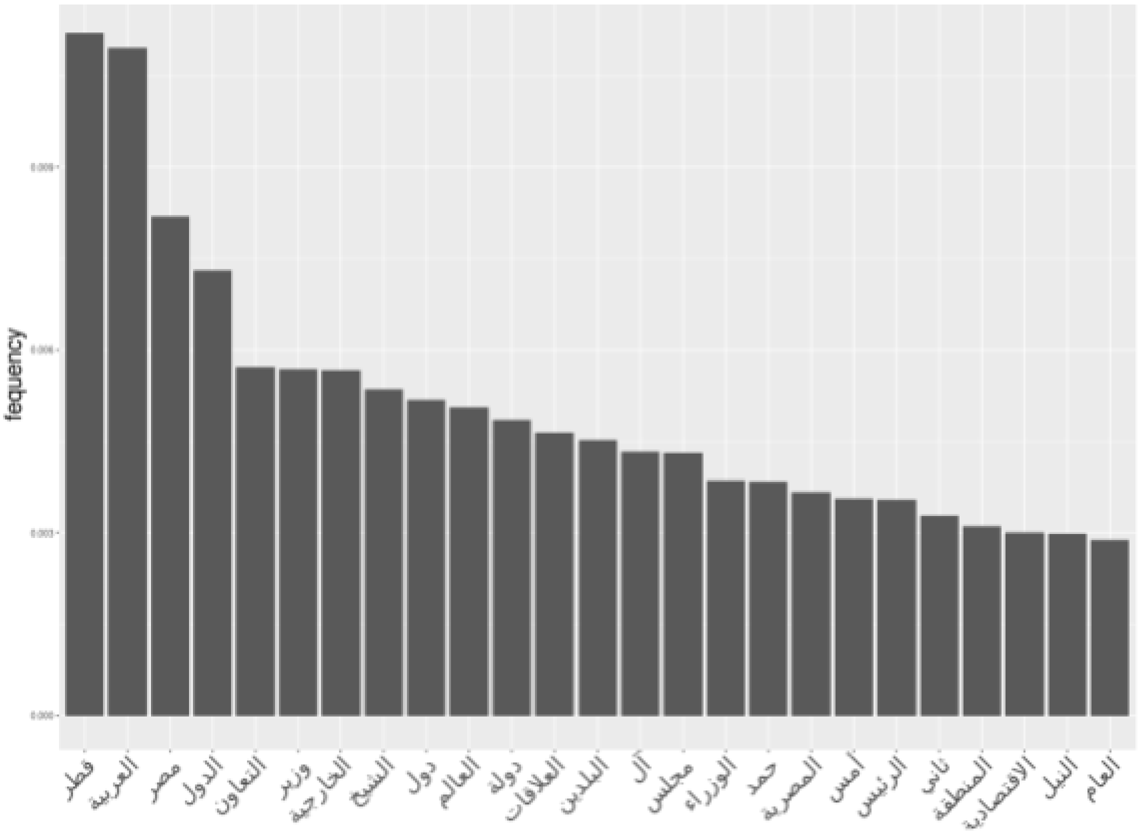


Figure 9. Most Frequent Words by Newspaper.
Military and domestic affairs.



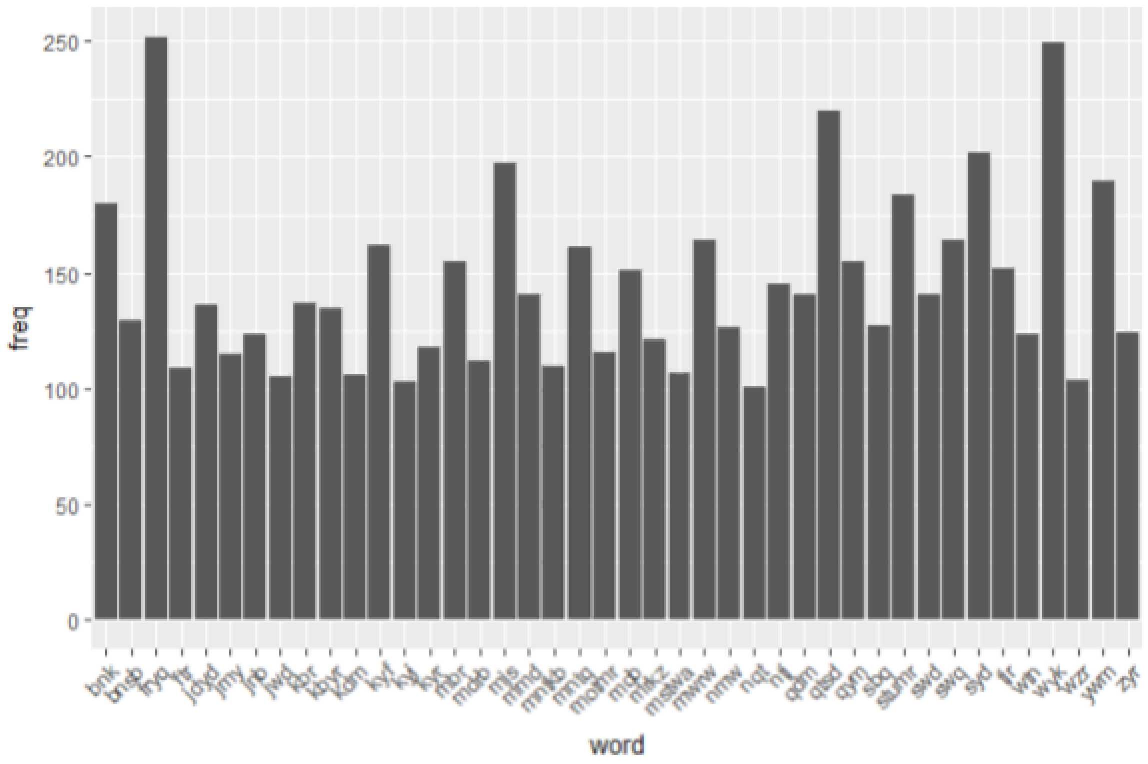


Figure 11. Transliterated histogram of
Al Watan MFWs.

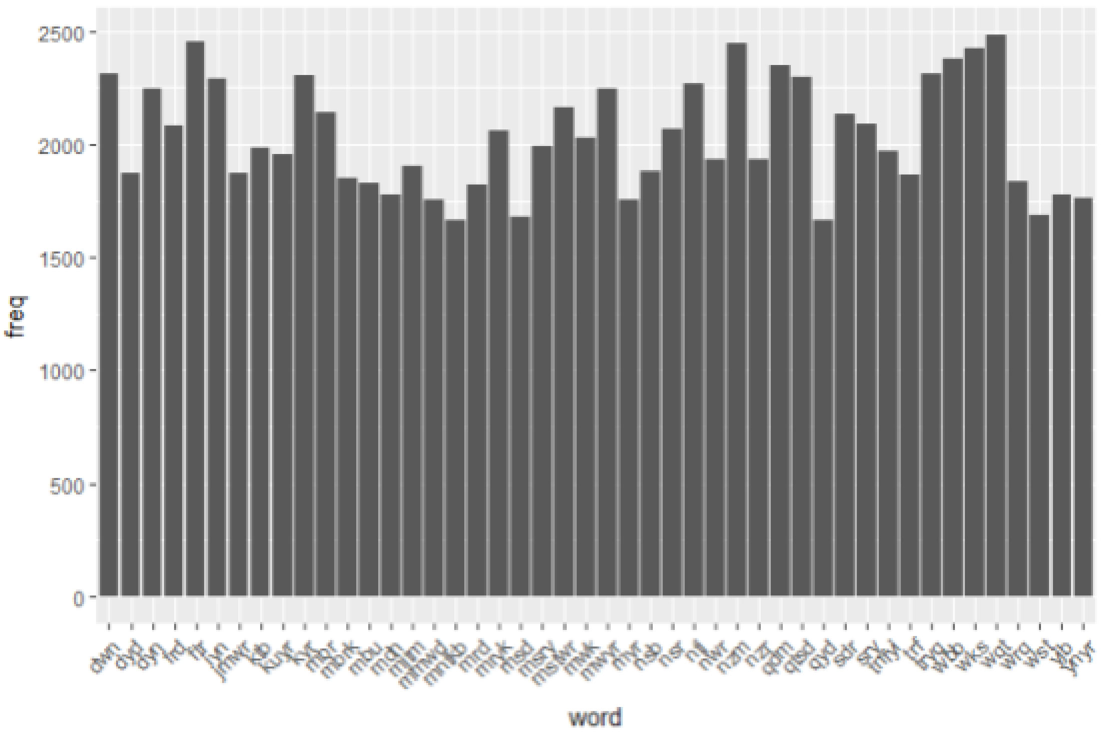


Figure 12: Transliterated histogram of
Ahram MFWs.

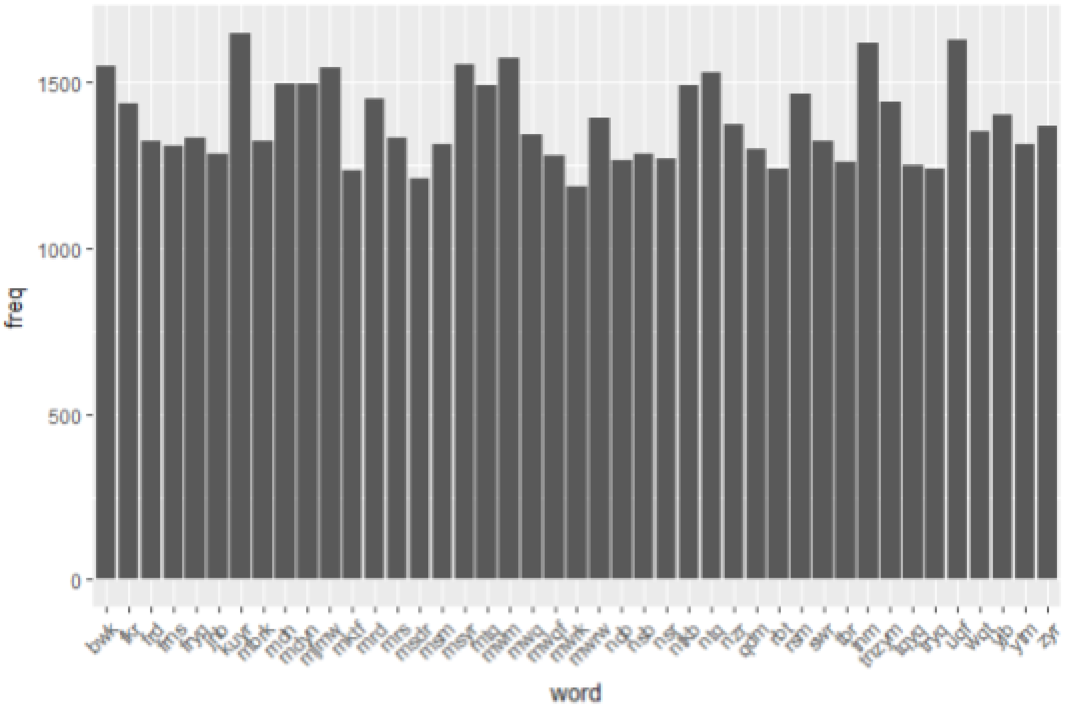


Figure 13. Transliterated histogram of
Thawra MFWs.

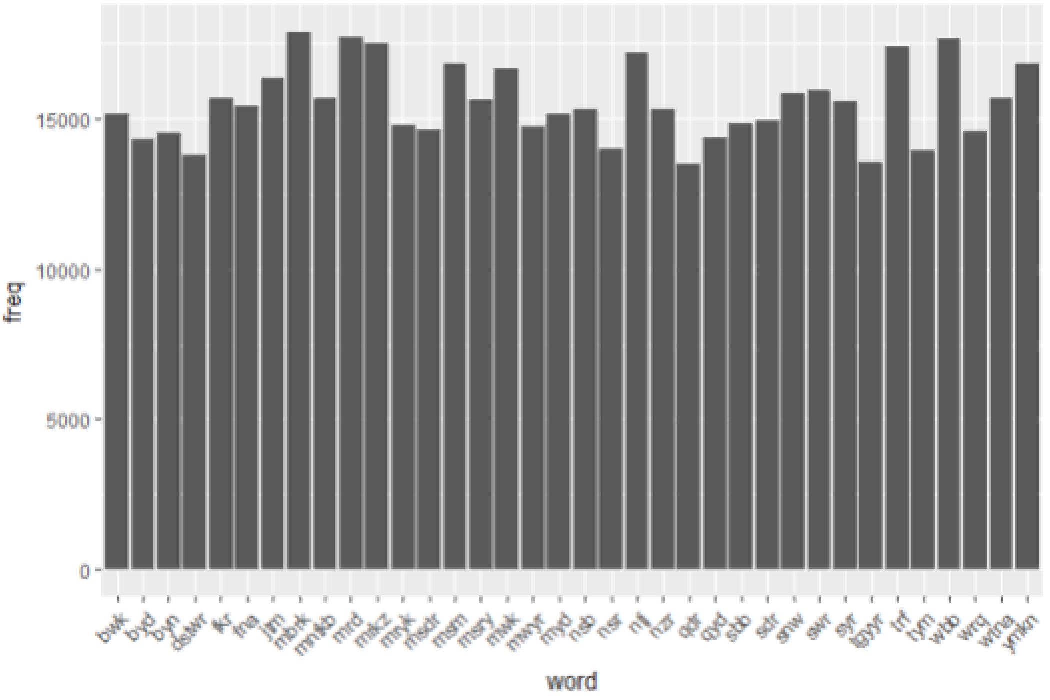
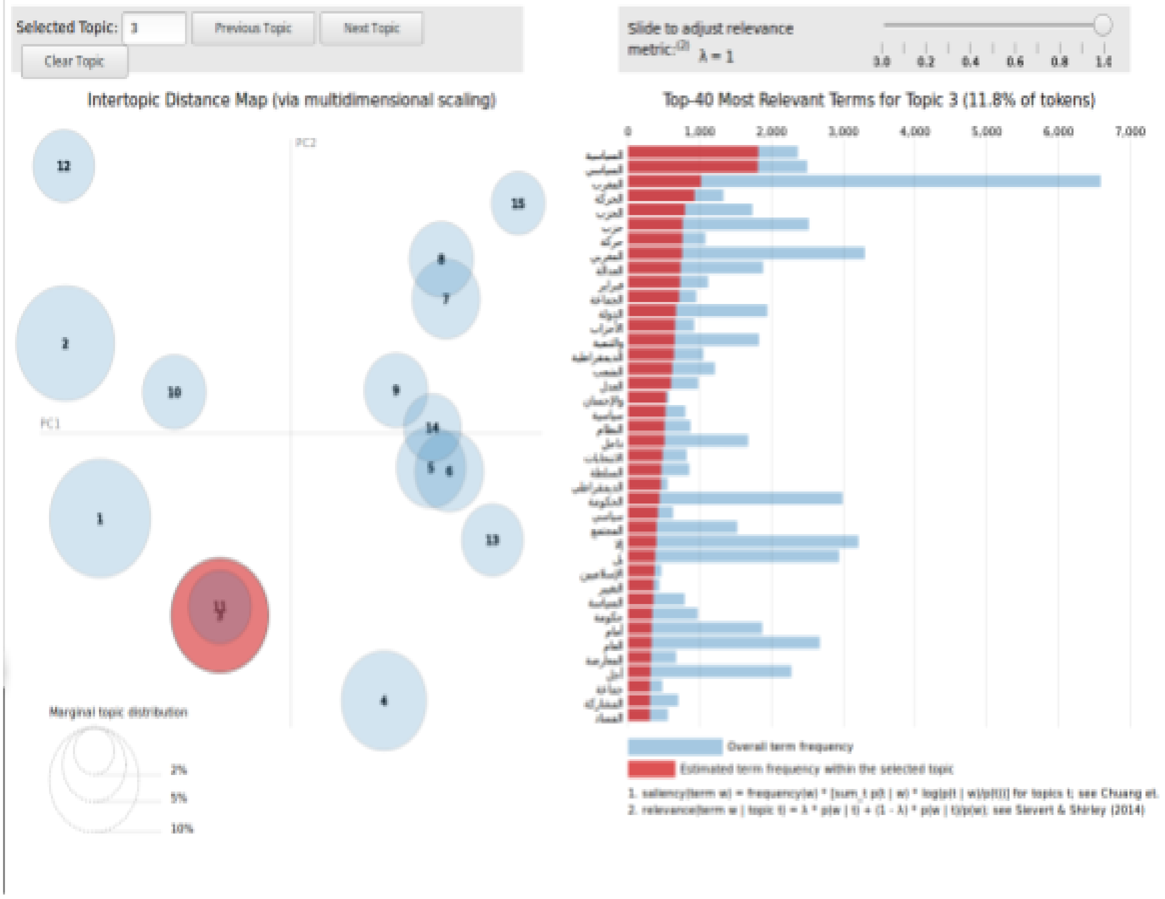


Figure 14. Transliterated histogram of
Al-Masri Al-Yawml MFWs.


Selected Topic: 3
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2) λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
12
15
8
7
2
10
9
14
5
6
1
13
3
4
Marginal topic distribution
2%
5%
10%
Top-40 Most Relevant Terms for Topic 3 (11.8% of tokens)
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
السياسة
السياسي
المغرب
الحركة
العرب
Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et.
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 15. Screenshot of created topics in LDAvis in the Hespress news source.

TABLE 3: Overview of possible news sources. From this informal database, we selected sites that could be easily scraped

Name of paper	**Country**	**Affiliation**
Hespress	Morocco	Independent
Al Arab	Qatar	Independent - pro-government
Al Raya	Qatar	Semi-official - pro-government
Al Rai	Kuwait	
al-Watan	Qatar	Independent
Al Sharq	Qatar	Pro-government but owned by private company
Al-Masri Al-Yawm	Egypt	Independent
Al-Thawra	Syria	
Al arab	Israel	
BBC Arabic	UK	Independent- Western source

Al-Ahram	Cairo, Egypt	
Al Jazeera	Doha	
al Watan	Northern Kuwait	Independent
Al Mustaqbal	Lebanon	
Al Bayan	UAE	
Al Ayam	Yemen	
Al Ghad	Jordan	
Al Tayyar	Lebanon	
Al Ittihad	UAE	
Al Khaleej	UAE	
Al Methaq	Yemen	
Azzaman	Iraq, International	
Al Madina	Saudi Arabia	