Arabic News Text Classification and Summarization

Final Defense Presentation

Tarek Kanan

June 4, 2015

Department of Computer Science, Virginia Tech

Committee: Drs. Edward Fox (Chair), Weiguo Fan, Roger Ehrich, Cliff Shaffer, Riyad Al-Shalabi

VirginiaTech

Outline

- Introduction (review): Overview, Research Questions, Hypotheses, Arabic Language
- Browsing Taxonomy
- P-Stemmer
- Text Classification
- RenA
- ALDA
- Template Summaries
- Conclusions

Overview of Research

- Text **Classification** for Arabic news articles using the generated taxonomy and applying a newly proposed stemmer called **P-Stemmer**

- Text **Summarization** for Arabic news articles after applying methods to extract the key points of the articles using different NLP tools, and filling in templates

Research Question

- The overall research question for this study is:
  - How and to what extent can we **classify** and **summarize** Arabic language text resources into Arabic text article categories and Arabic readable summaries without direct human interaction while achieving **high quality results**?

Research Questions

- From the above question, we can derive more specific research questions:
  - How can we create a simple, but general, classification **taxonomy** for Arabic news articles?
  - What is the most effective approach for **classifying** Arabic news articles that leads to high quality labeling?
  - How can **stemming** enhance Arabic text classification?
  - How can we apply natural language processing methods to create good **summaries** of Arabic news articles?
  - Are the produced summaries as **good** as human summaries?

Main Hypothesis 1

- The quality and accuracy of Arabic text **classification** using the proposed **taxonomy** and **stemmer** is better than with state-of-the-art approaches and systems.
- Hypothesis 1-1:
  - The proposed **taxonomy** is easy to use, works well with any **Arabic newspaper**, and is compatible with the **IPTC** system.
- Hypothesis 1-2:
  - The proposed **combined stemming and text classification method** is **more effective** than state-of-the-art pairs of stemmers and text classification methods.

Main Hypothesis 2

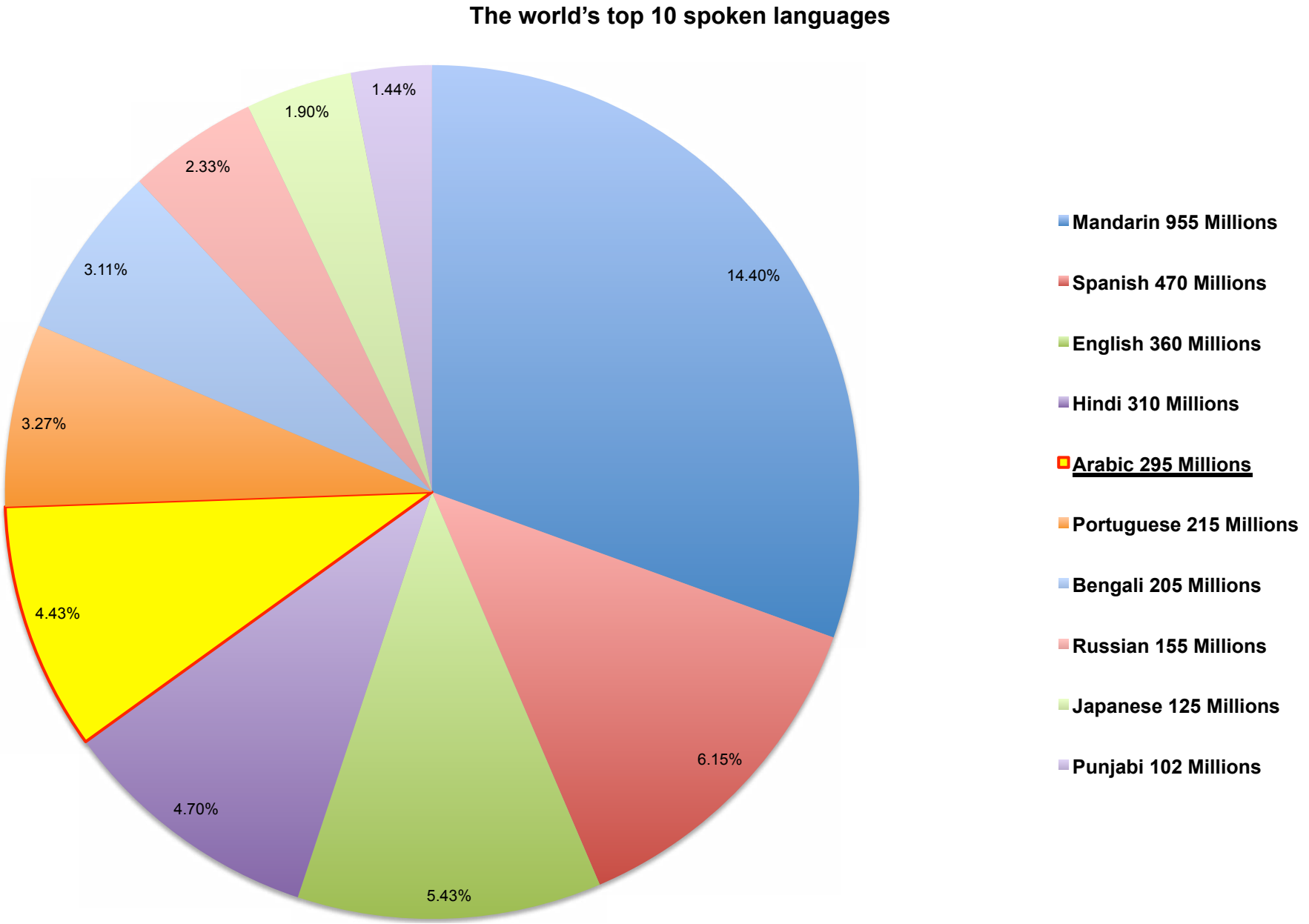
- The proposed automatic Arabic text **summarization** approach, applied to news articles, will give accurate summaries that are relevant to the news articles.

- Hypothesis 2-1:
  - The proposed summarization approach will produce **high quality** Arabic news article summaries, by using text **extraction** methods to fill in a developed **template,** evaluated through human assessment.

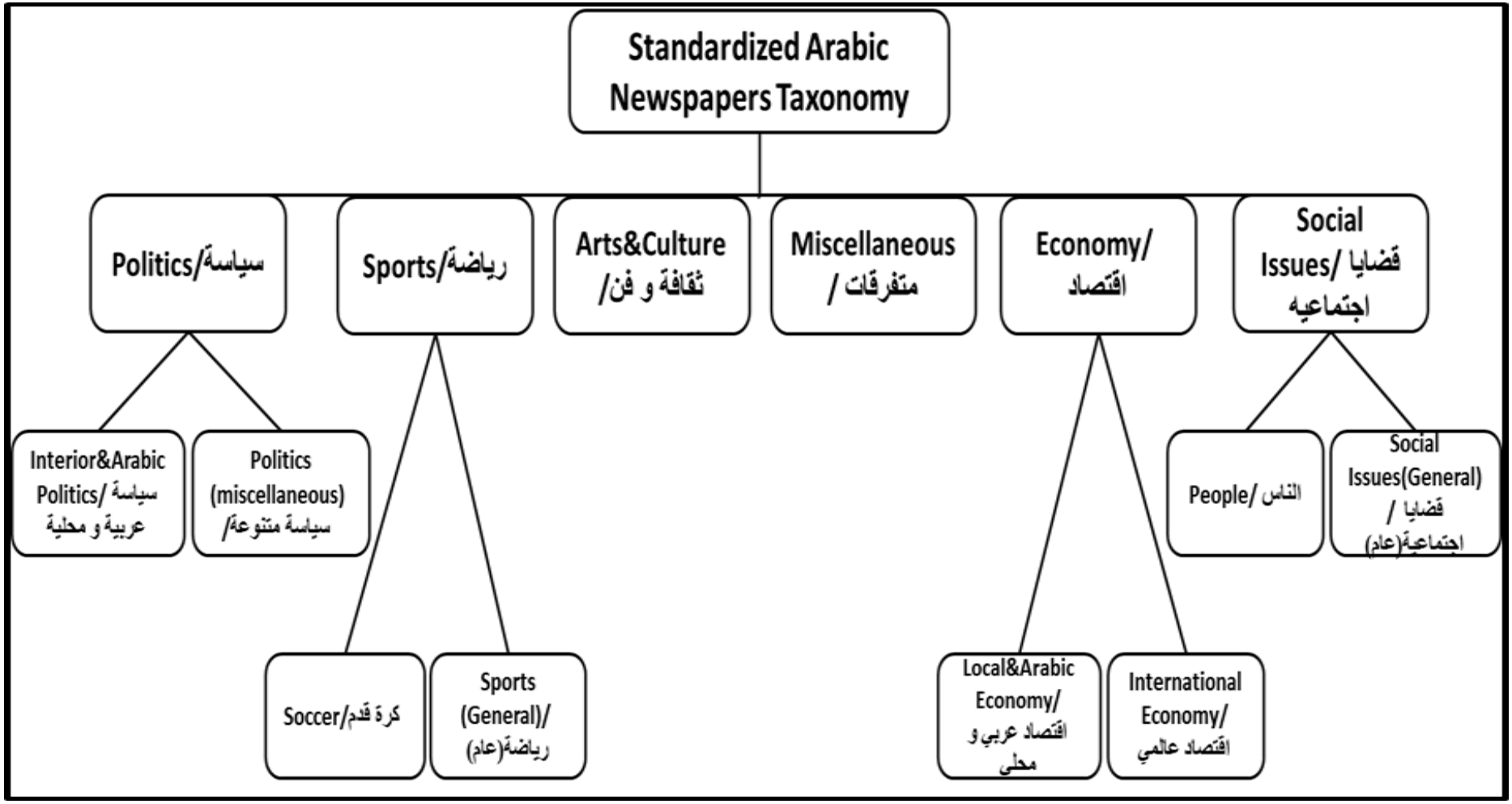
The Arabic Language

- Arabic is a widely used global language that has major **differences** from the most popular, e.g., English and Chinese.
- The Arabic language has many **grammatical** forms, varieties of word **synonyms**, and different word meanings that vary depending on factors like:
  - Word order
  - **Diacritics**

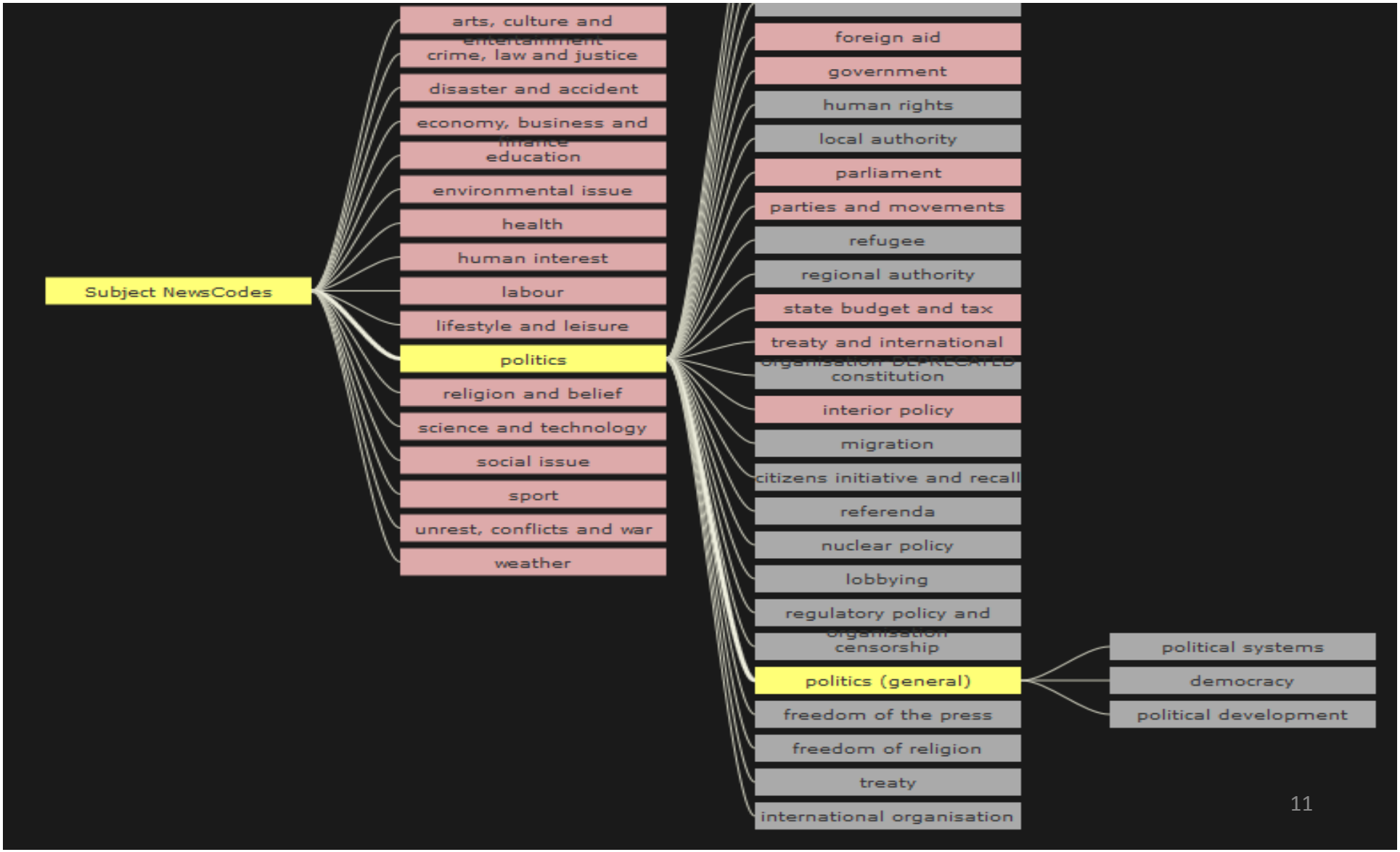
The Arabic Language



Standardized Taxonomy



Building a Standardized Categorization System for Arabic Newspapers: Subject NewsCodes in the IPTC Taxonomy



What is Stemming

- The process for reducing inflected or derived words to their word stem, base, or root form
- Two types for Arabic stems:
  - Root: the goal of a **root**-based stemmer is to extract the very basic form for any given word.
  - Light: the goal of a **light** stemmer is to find the canonical form of an Arabic word by removing prefixes and/or suffixes

P-Stemmer

- Called Prefix Stemmer (**P-Stemmer**)
- It is a modified version of Larkey's **light10** stemmer
  - Larkey's stemmers are **popular** Arabic light stemmers
  - **Larkey's** five versions of light stemmers:
    - Light1, Light2, Light3, Light5, and **Light10**
- P-Stemmer only removes **prefixes**

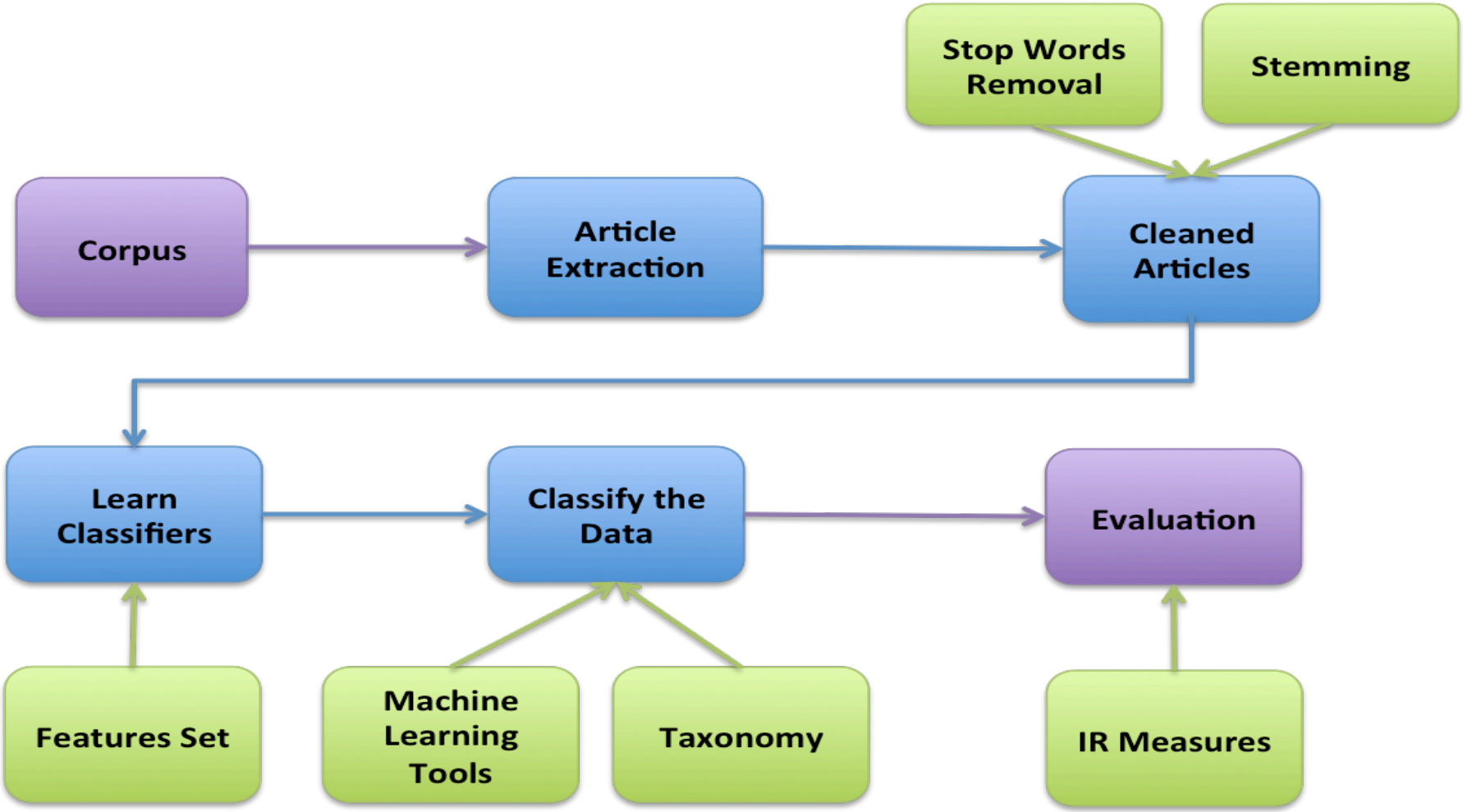
P-Stemmer Examples

Word	Light10	P-Stemmer
كالصادرات As the imports	صادر Took	صادرات Imports
والوحدات And the units	وحد Aggregate	وحدات Units
المكتبات The libraries	مكتب Office	مكتبات The library
المباحثات The talks	مباحث Investigation	مباحثات Talks

P-Stemmer

- Available from:
- https://github.com/tarekll/P-Stemmer

Arabic Text Classification



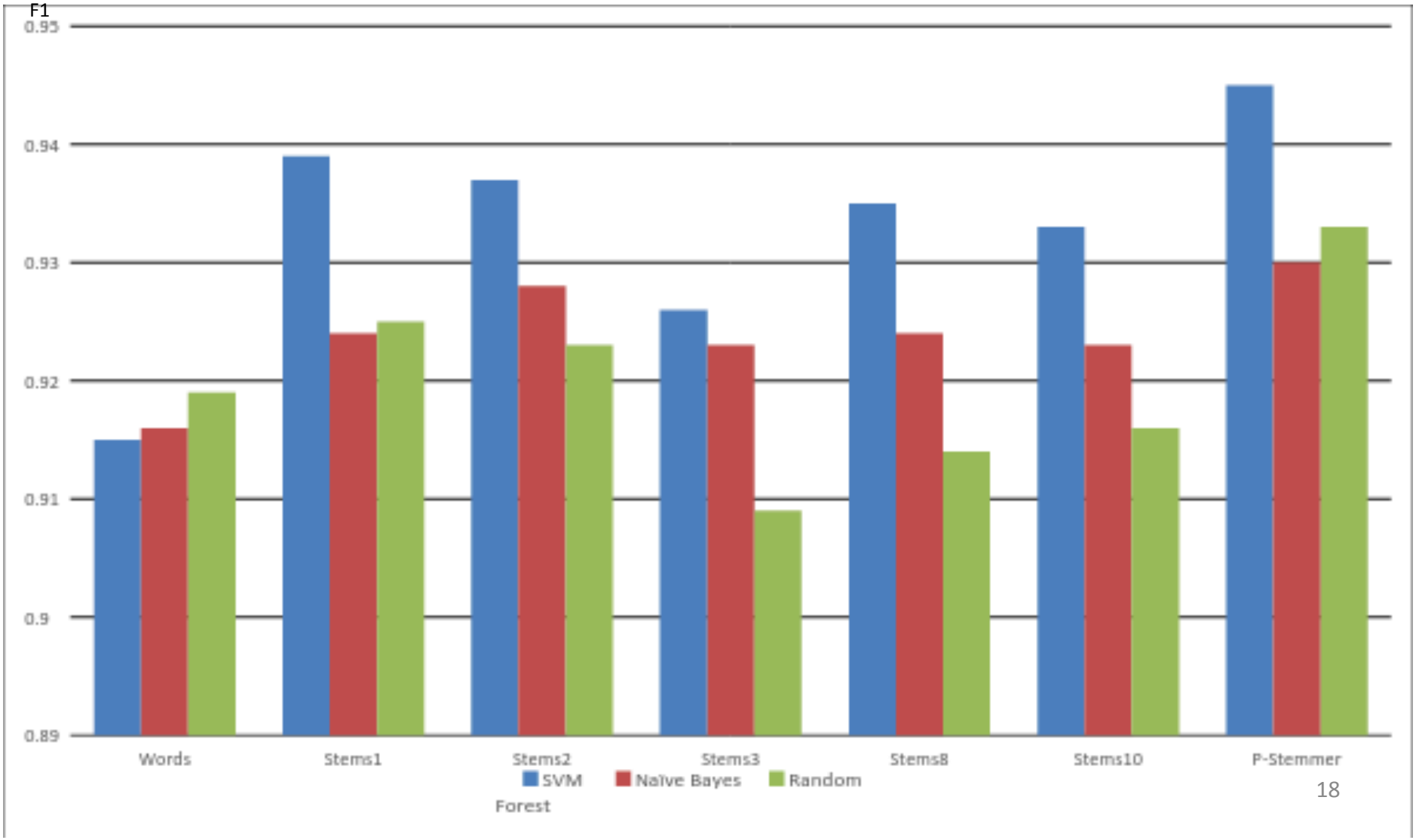
Arabic Text Classification

- We used **SVM, NB**, and **RF** classifiers to
  - Judge the performance of P-Stemmer
  - Compare it with the other listed approaches

We categorized the data into one of five main categories

- Sports
- Economics
- Politics
- Art & Culture
- Social Issues

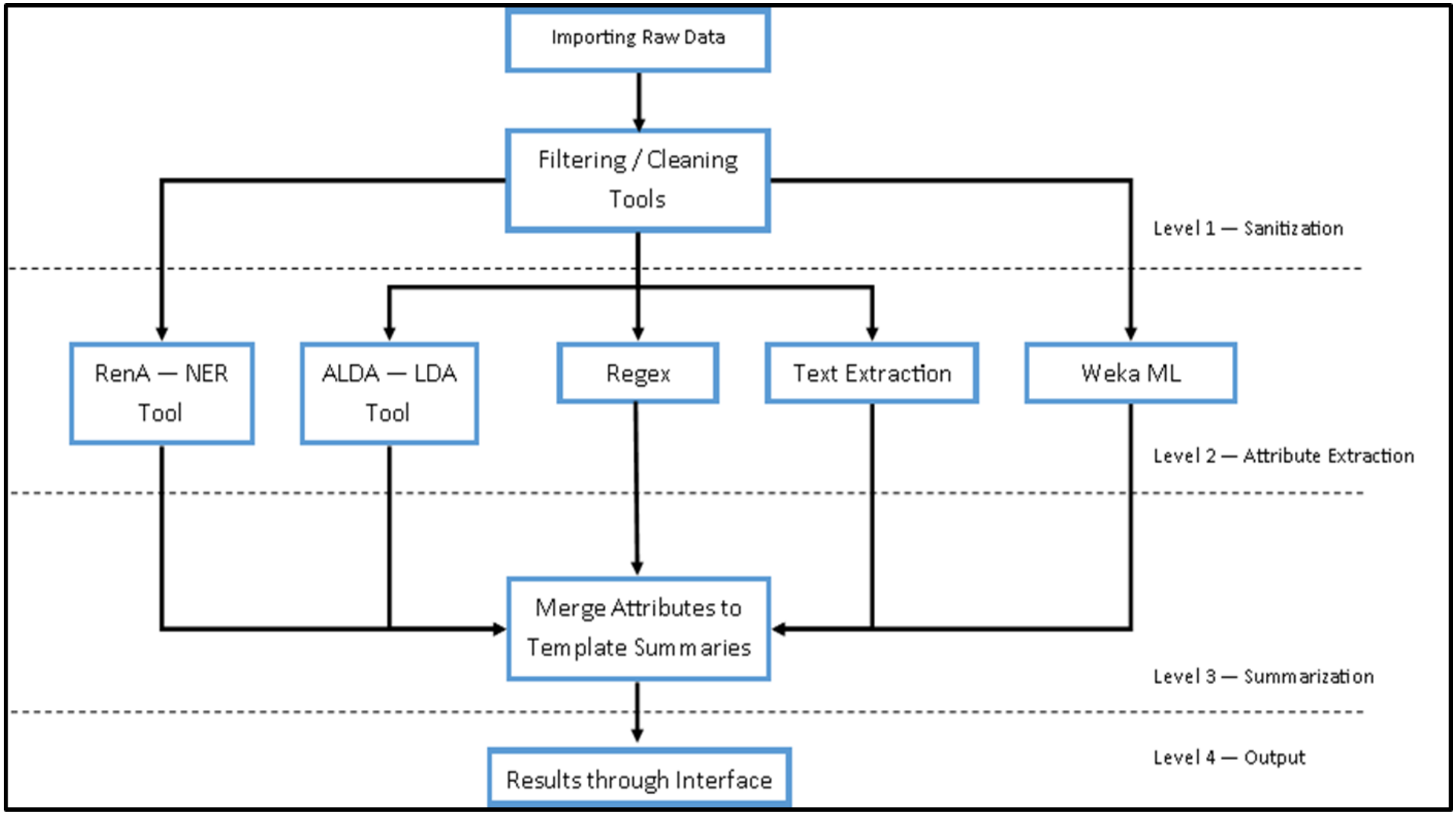
Arabic Text Classification Evaluation



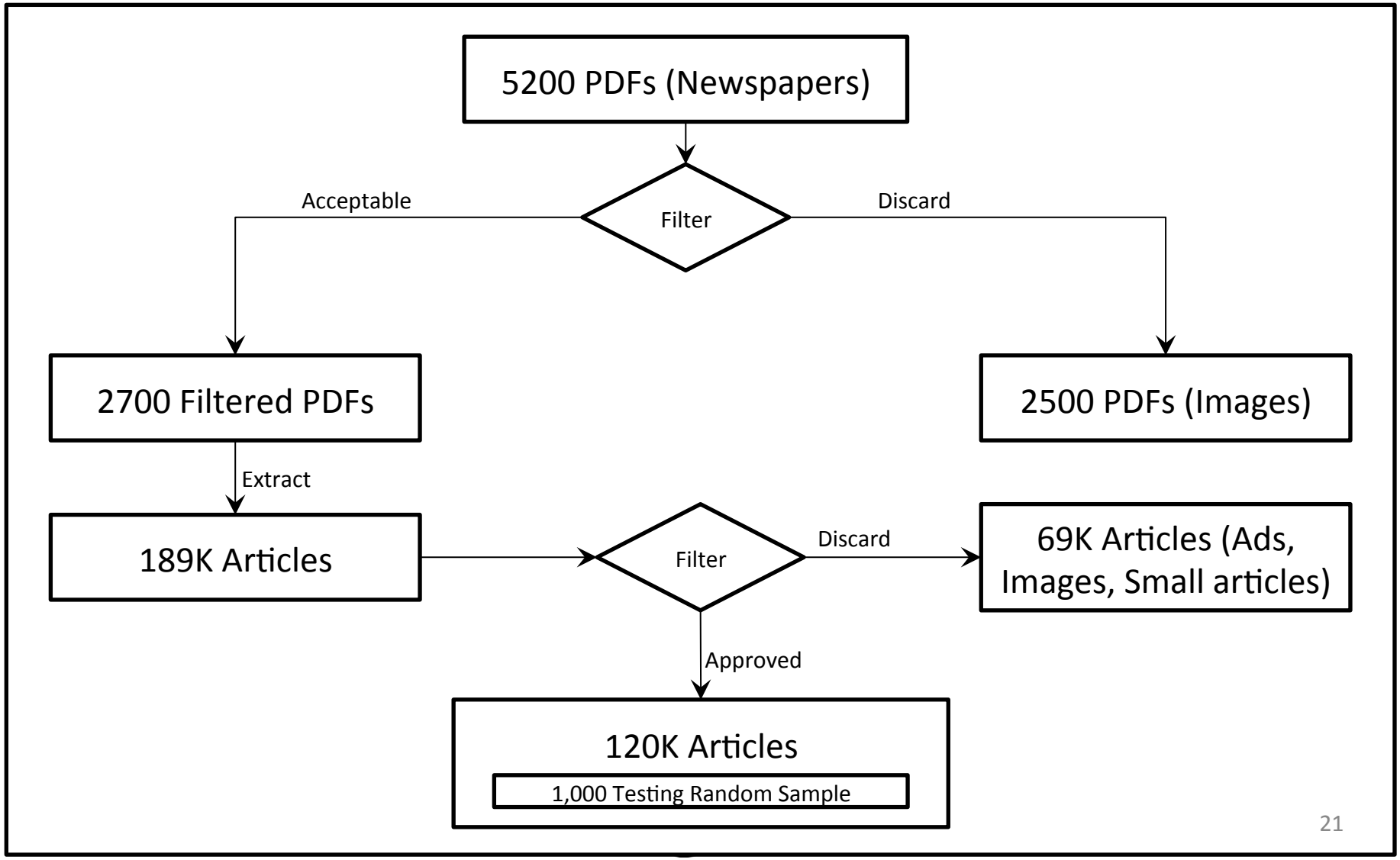
Significance Test

- We used the Wilcoxon signed-ranked test to compare our proposed stemmer and each of Larkey's 5 stemmers.
- Null hypothesis: "The median difference of the F1 measure of P-Stemmer and each one of Larkey's stemmers (Stem1, Stem2, Stem3, Stem8, or Stem10) is less than or equal to zero"
- We successfully rejected our null hypothesis, for each one of the five tests.
- We concluded that, using the F1 measure for evaluation, our P-Stemmer is statistically significantly better than each one of the five Larkey's stemmers.

Summarization Architecture Diagram



Dataset Preparation



NER

- Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction
- It seeks to locate and classify elements in text into pre-defined categories such as:
  - The names of persons, organizations, locations, expressions of times, dates, etc.

NER: Results (English)

AFP/Madrid
Zinedine Zidane is shaping up as a future coach of Real Madrid, present incumbent Carlo Ancelotti said yesterday.
Zidane, who is currently coaching the Real reserve side Castilla, "has all the qualities" required to take the helm of the club, Ancelotti told a news conference. "I enjoy Zidane's work, he's doing very well," Ancelotti said.
After a difficult start of the season, Castilla are top of Spain's third tier league. "He's doing very well in his first year in charge. He's taken Castilla to first place and he needs to keep up the good work.
"It's pretty clear to me he has all the qualities to coach a big team. And that includes Real Madrid," said the Italian manager, who appointed the French legend last season.
After seeing Castilla loses five of their first six initial games, Zidane has turned things around and his young charges have now lost just once in the past four months.
They could increase their lead when they take on Athletic Bilbao's reserves on Sunday, a match which could see Norwegian teenage prodigy Martin Odegaard, snapped up from under the noses of many European giants in the transfer window, could make his debut.

Person: Zinedine, Zidane, Carlo, Ancelotti, Martin, Odegaard
Organization: Real, Madrid, Castilla, Athletic, Bilbao
Location: Spain, Madrid, Bilbao, Norway, Europe

RenA: Results (Arabic)

الدوحه: تنظم جمعيه الهندسه والتكنولوجيا مساء ٣١ من نوفمبر الجاري لقاء للمهندسين المقيمين والزائرين بكلي شمال الاطلنطي لتبادل الخبرات والتعرف على بعض الابتكارات التي تحدث في قطر.وتعتبر جمعيه الهندسه والتكنولوجيا اكبر جمعيه حرفيه للمهندسين في اوروبا وتضم اكثر من ٠٥١,٠٠٠ عضو في ٧٢١ دوله، وسيجتمع بعض اعضائها المقيمين في قطر مع عدد من المهندسين والعلماء وطلبه الجامعات. وقال ماكس رينو: »هذه فرصه للمهندسين هنا لتبادل الخبرات مع اترابهم المحترفين وللترويج للابتكارات التي تتحقق في قطر«.فيما قال انطوني بيكر المتحدث باسم اللجنه المنظمه: »نود ان نتقدم بالشكر لكليه شمال الاطلنطي لاستضافه ودعم هذا الحدث، ونامل ان يشجع هذا الحدث الشباب على الدراسه والتفكير في فرص العمل المثيره والمجديه في مجالات العلوم والتكنولوجيا والهندسه

Person: ماكس، رينو، انطوني، بيكر

Organization: جمعيه الهندسه والتكنولوجيا، كليه شمال الاطلنطي

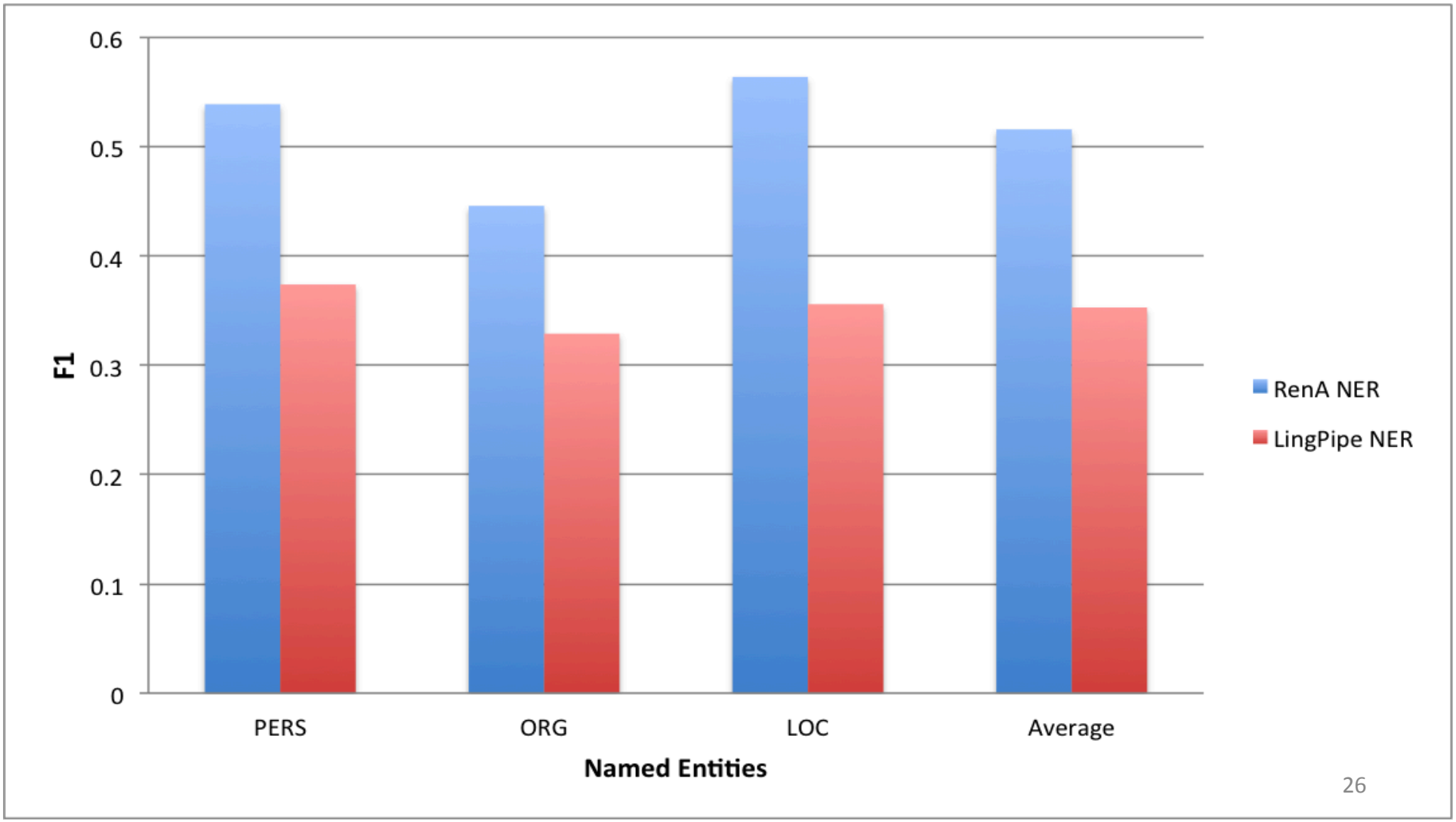
Location: الدوحه، قطر،اوروبا

VirginiaTech

Baseline Corpus

- Could not find labeled Arabic news article corpus to:
  - Test and evaluate the Named-Entity Recognition (NER) results
  - Compare our NER with existing NERs
- Decided to **build** our **baseline** corpus from our dataset
- 1000 articles, random sample
- 10 participants, each:
  - Assigned equal number of articles
  - Extracted the 3 types of named entities (Person, Organization, and Location)
- Extracted entities checked twice

RenA: Evaluation



RenA: Evaluation

	RenA NER			LingPipe Toolkit NER		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
PERSON	0.826	0.497	0.539	0.582	0.371	0.374
ORGANIZATION	0.813	0.421	0.446	0.39	0.377	0.329
LOCATION	0.77	0.558	0.564	0.55	0.338	0.356
Average	0.803	0.492	0.516	0.507	0.362	0.353

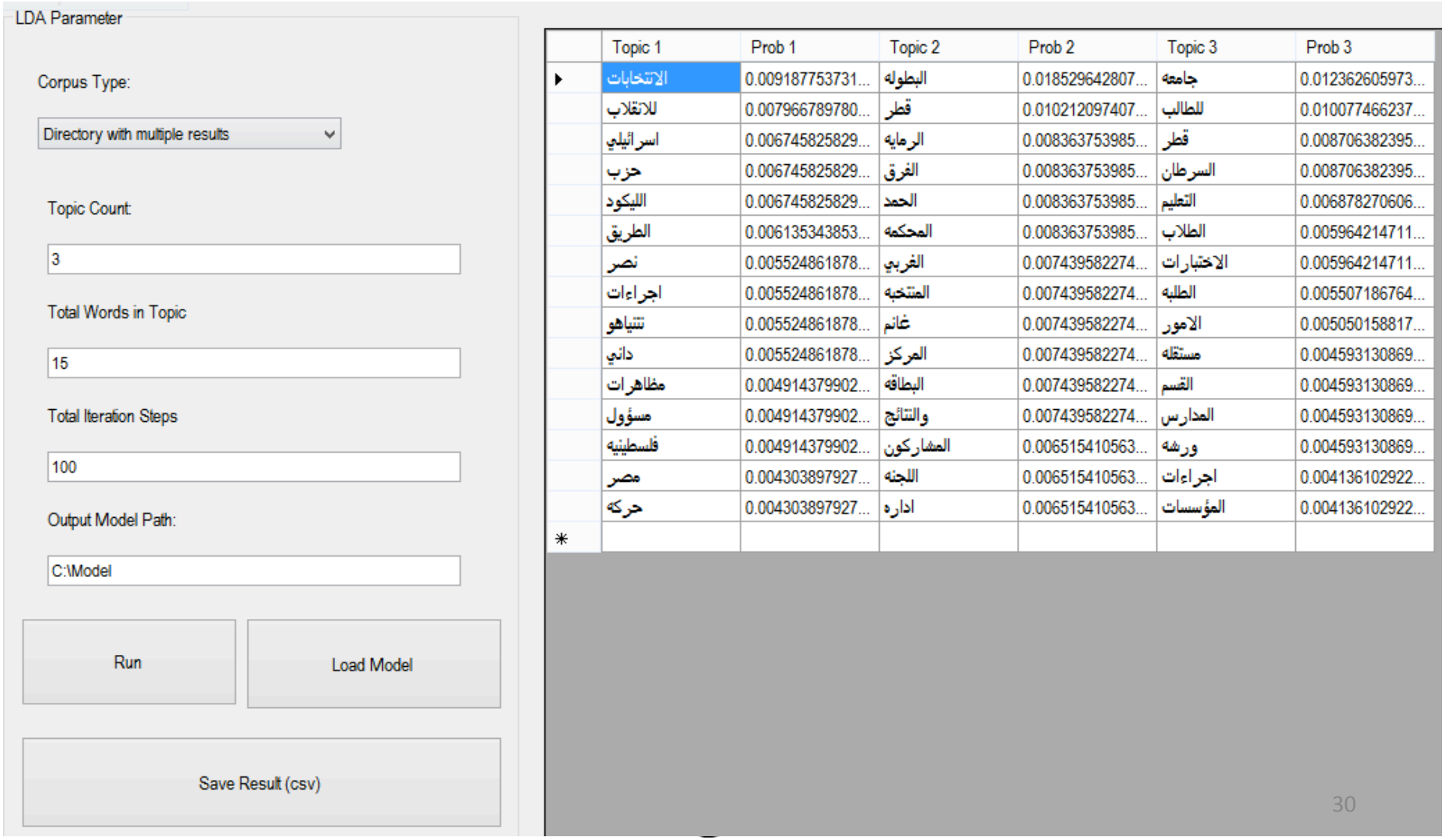
RenA

- Available from:
- https://github.com/tarekll/RenA

Topic Identification

- It's the way to identify what are the topic(s) in a set of documents
- Given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently
- LDA, one of the popular topic modeling algorithms

ALDA: Screen Shot



	Topic 1	Prob 1	Topic 2	Prob 2	Topic 3	Prob 3
▶	الانتخابات	0.009187753731...	البطوله	0.018529642807...	جامعه	0.012362605973...
	للانقلاب	0.007966789780...	قطر	0.010212097407...	للطالب	0.010077466237...
	اسرائيلي	0.006745825829...	الرمايه	0.008363753985...	قطر	0.008706382395...
	حزب	0.006745825829...	الغرق	0.008363753985...	السرطان	0.008706382395...
	الليكود	0.006745825829...	الحمد	0.008363753985...	التعليم	0.006878270606...
	الطريق	0.006135343853...	المحكمه	0.008363753985...	الطلاب	0.005964214711...
	نصر	0.005524861878...	الغربي	0.007439582274...	الاختبارات	0.005964214711...
	اجراءات	0.005524861878...	المنتخبه	0.007439582274...	الطلبه	0.005507186764...
	نتنياهو	0.005524861878...	غانم	0.007439582274...	الامور	0.005050158817...
	داني	0.005524861878...	المركز	0.007439582274...	مستقله	0.004593130869...
	مظاهرات	0.004914379902...	البطاقه	0.007439582274...	القسم	0.004593130869...
	مسؤول	0.004914379902...	والنتائج	0.007439582274...	المدارس	0.004593130869...
	فلسطينيه	0.004914379902...	المشاركون	0.006515410563...	ورشه	0.004593130869...
	مصر	0.004303897927...	اللجنه	0.006515410563...	اجراءات	0.004136102922...
	حركه	0.004303897927...	اداره	0.006515410563...	المؤسسات	0.004136102922...
*						

ALDA: Article/Topic (Arabic)

طرابلس – رويترز: قال مسؤول ان رجال قبائل ليبيين انهوا حصارهم لحقل الشراره النفطي لكن لا يتسنى استئناف الانتاج لحين انتهاء احتجاج منفصل عند خط انابيب مرتبط بالحقل. وكان قبليون وحراس امن اغلقوا الحقل ال ذي تبلغ طاقته ٠٤٣ الف برميل يوميا بجنوب البلاد في فبراير شباط للضغط من اجل مطالب ماليه وسياسيه وهو ما زاد من حده الحصار المفروض على موانئ نفط في الشرق. وقال حسن الصديق مدير حقل الشراره لرويترز ان المحتجين الذين اغلقوا الحقل تركوا المكان لكن لا يمكن استئناف العمل به نظرا لان الصمامات ما زالت مغلقه. واضاف ان هناك مفاوضات تهدف الى انهاء اغلاق صمامات خط الانابيب في الجبال الغربيه ويامل المهندسون باستئناف الضخ في غضون اسبوع. واغلقت مجموعه اخرى من المحتجين في منطقه ال زن تان في الغرب خطوط الانابيب من اجل مطالب ماليه وسياسيه. واغلق محتجون حقل الشراره اكثر من مره.وكان انتاج ليبيا يبلغ نحو ٤،١ مليون برميل يوميا حتى منتصف عام ٣١٠٢ حين بدات الاحتجاجات التي قلصته الى اكثر قليلا من ٠٠٢ الف برميل يوميا.حقل الشراره الليبي ما زال مغلقا رغم انتهاء الاحتجاج بكتيريا ماصه للغازات الطبيعيه لمواجهه التسرب النفطي

Probability – Topic
0.0361768646717284, الشراره
0.027244305493523 9, النفطي
0.027244305493523 9, احتجاج
0.027244305493523 9, انابيب
0.027244305493523 9, برميل
0.027244305493523 9, المحتجين
0.018311746315319 3, حصارهم
0.018311746315319 3, استئناف
0.018311746315319 3, الانتاج
0.018311746315319 3, انتهاء

ALDA: Article/Topic (English)

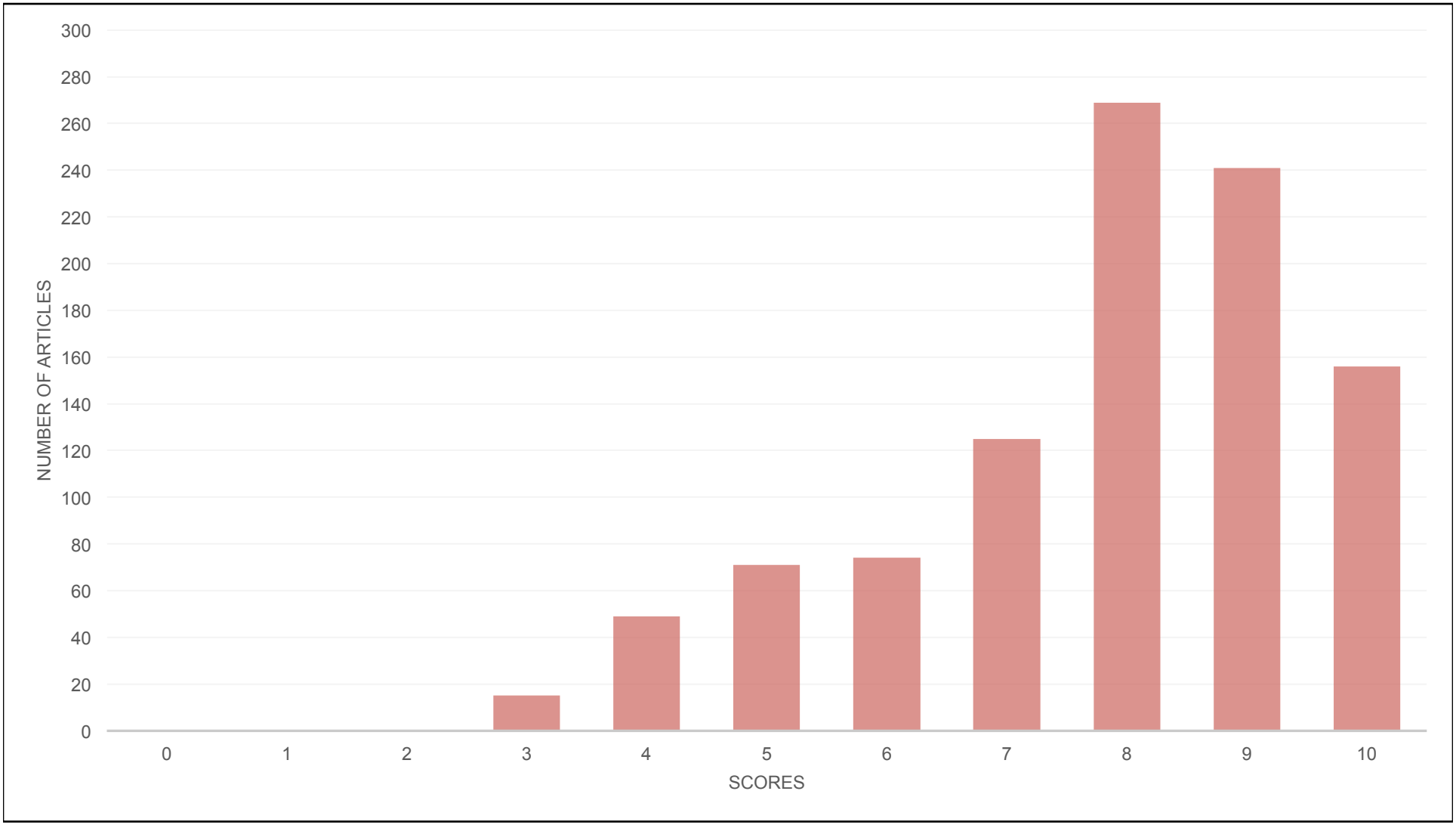
Tripoli - Routers: An official said the tribesmen from Libya ended their closure of the oil field of AlSharara, but it is not possible to resume production until the end of a separate protest connected to the field pipelines. The security guards blocked a field that has a capacity of 34 thousand barrels per day south of the country in the month of February to lobby for financial and political demands which increased the severity of the siege imposed on the oil. Hasan Alsadeq, AlSharara oil field director, said to Routers that the protesters left the field but can not resume work and that he hopes to resume work within a week. Closing the filed happened more than once. Libya's oil production was 4.1 million barrels per day.

- **AlSharara, Oil, Protest, Pipelines, Barrel, Protestors, Siege, Resume, Production, Ends**

ALDA: Evaluation

- 10 participants; each received 100 articles and their corresponding topics from the 1000 random sample
- Participants asked to evaluate the relevance of the topics
- Each topic/article pair evaluated twice, then averaged
- Count the frequencies of each rating

ALDA: Evaluation Results



VirginiaTech

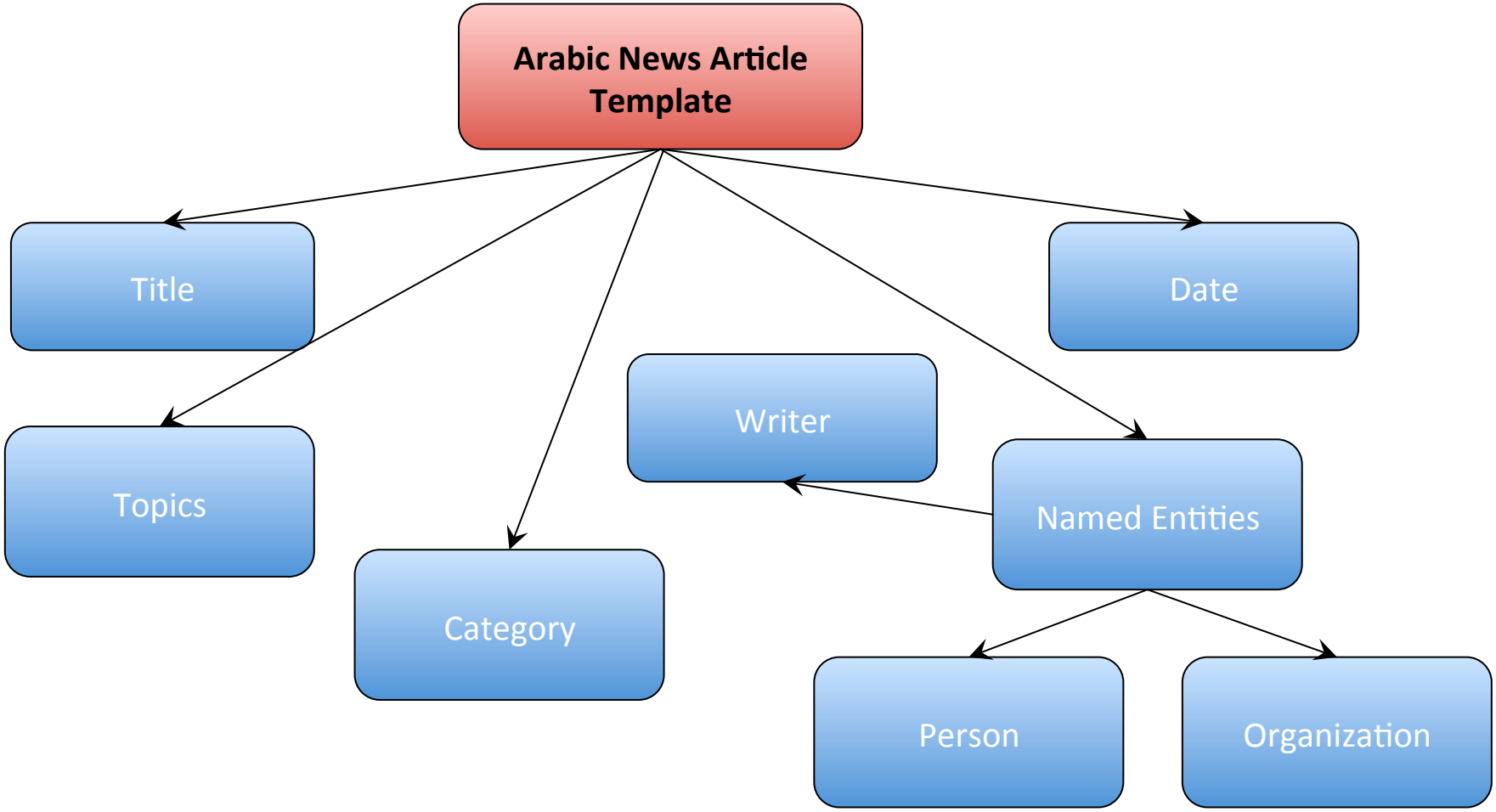
ALDA: Evaluation Results

Rating	0	1	2	3	4	5	6	7	8	9	10
Frequencies	0	0	0	15	49	71	74	125	269	241	156

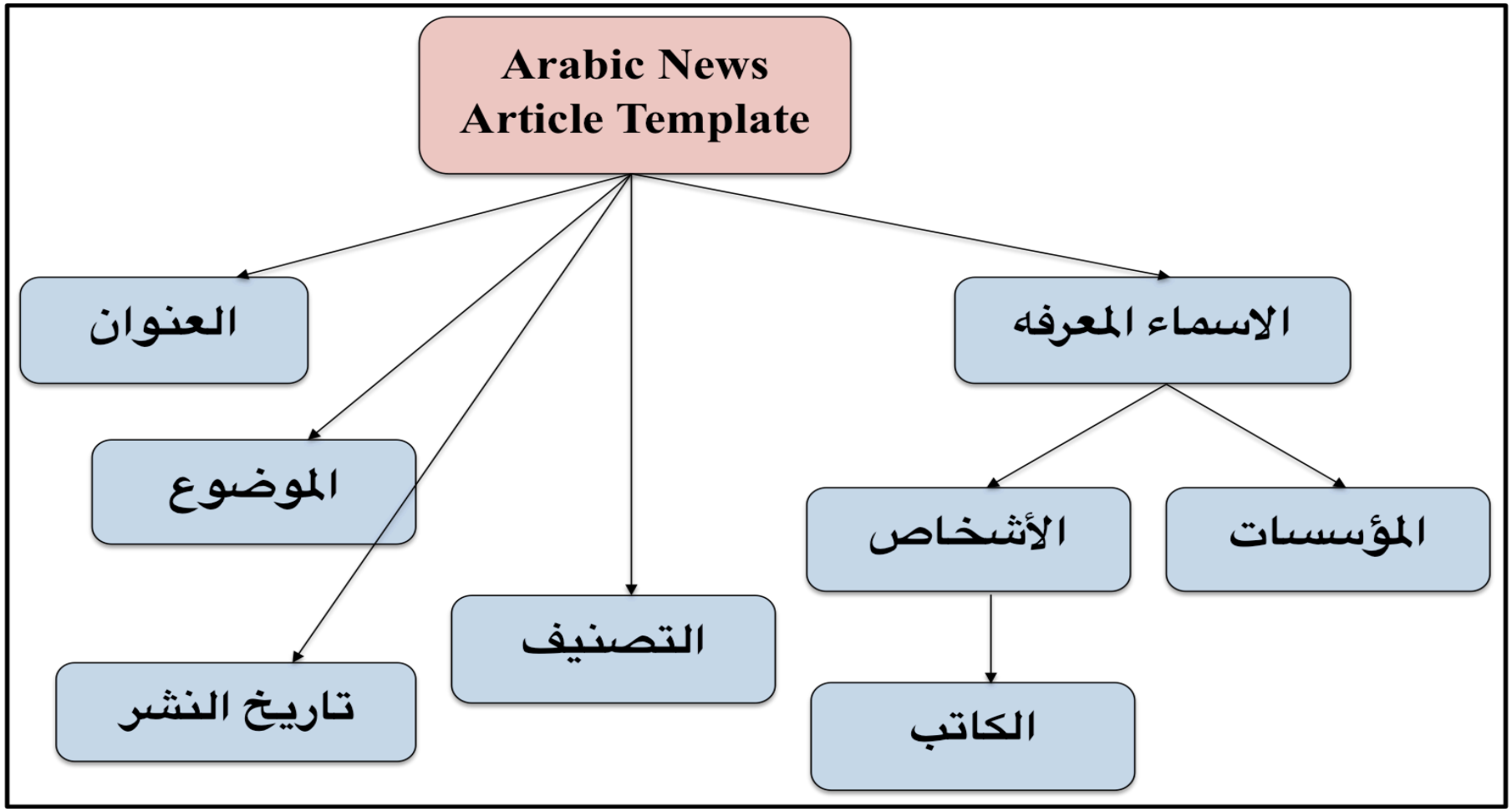
ALDA

- Available from:
- https://github.com/tarekll/ALDA

Summary Template's Attributes (English)



Summary Template's Attributes (Arabic)



Template Summaries Description

Template Attribute	Description
Writer	The first Person named entity extracted using NER
Date	The publishing Date extracted using regular expressions
Title	The article title, probably the first line in the article
Person(s)	The Person(s) named entity(ies) extracted using NER
Organization(s)	The Organization(s) named entity(ies) extracted using NER
Topic	The main Topic in the article generated using Arabic LDA tool
Category	The Category of the article generated using a classification algorithm

Template (Arabic/English)

{Title} : {العنوان}

{Publication Date} : {تاريخ النشر}

{Writer} : {الكاتب}

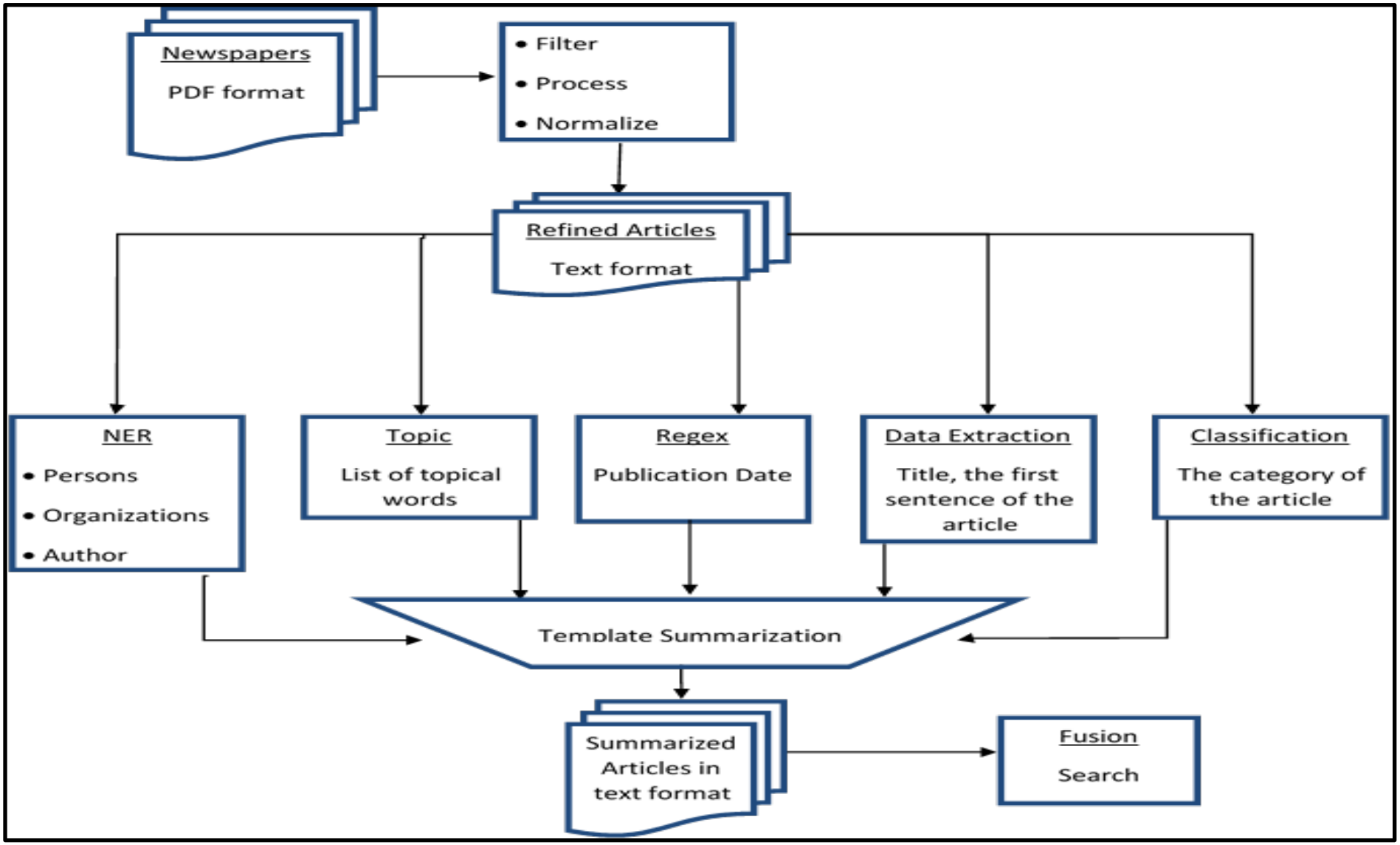
{People Mentioned} : {الأشخاص المشار اليهم}

{Organizations Mentioned} : {المؤسسات المشار إليها}

{General Topical Category} : {التصنيف العام}

{Words in Main Topic} : {الكلمات في الموضوع الرئيسي}

Overall Dataflow Diagram



Fusion

- The final results of this work are incorporated into one of the collections of the ELISQ project that has been indexed and made available to search through LucidWorks Fusion
- For more details see:
  - http://10.100.121.44:8000
    - Choose “Arabic News Articles Template Summaries” collection

Arabic News Article Example

الدوحه 25/8/2012- انور الخطيب:

تحقيق الامن والاستقرار واعاده بناء مصر من جديد على اساس المساواه والعداله الاجتماعيه. مهام ملحه ارتاى عدد من الكتاب والمحللين السياسيين في قطر ان على الدكتور محمد مرسي رئيس مصر المنتخب ان يركز عليها في بدايات عهده الجديد.. مؤكدين ان نجاح الدكتور مرسي كاول رئيس مصري منتخب من الشعب يؤسس لبناء مصر الجديده وسينعكس ايجاب ا على باقي الدول العربيه. ونوهوا بما وصفوه ب »الحمل الثقيل« الذي ورثه مرسي، الذي لن يتمكن دون تكاتف كل القوى السياسيه في مصر، من بناء مصر الجديده على قواعد جديده تضمن لجميع المواطنين المساواه في الفرص وتحقيق العداله الاجتماعيه واعاده عجله التنميه الى الدوران في البلاد.

فمن جانبه راى الدكتور عبد الحميد الانصاري عميد كليه الشريعه والقانون السابق في جامعه قطر ان المهمه الاساس للرئيس الجديد تحقيق الاصلاح وتنفيذ الوعود التي قطعها خلال حملته الانتخابيه. مضيفا ان تحقيق الاستقرار في مصر سينعكس ايجاب على الوضع المصري الداخلي وبالضروره ان ذلك سينعكس على الدول العربيه. واكد الدكتور الانصاري ان انتخاب الرئيس الجديد تم بصوره ديمقراطيه وان صندوق الاقتراع هو من حسم هويه الفائز وان على جميع المصريين ان يتقبلوا هذه النتيجه ويتقبلوا نتيجه صندوق الاقتراع. فالفائز اصبح رئيسا لجميع المصريين مهما كانت انتماءاتهم السياسيه، وفي المقابل فان على الرئيس الفائز والذي كان مرشحا لجماعه الاخوان المسلمين ولحزب الحريه والعداله ان يمارس مهامه كرئيس لمصر وليس رئيسا لحزب والا يفرق بين ابناء الوطن حسب انتماءاتهم السياسيه او الدينيه. وكرر التاكيد على ضروره ان يبادر الرئيس الجديد بتحقيق الاصلاح الداخلي والتركيز على قضيه الاقتصاد الذي وصل في مصر الى مرحله الحضيض وان يعمل على دفع المصريين لمزيد من الانتاج والعمل لانقاذ اقتصاد البلاد المتهاوي. من جهتها عبرت الكاتبه الدكتوره موزه المالكي عن سعادتها بنجاح الثوره المصريه ونجاح اول تجربه ديمقراطيه حقيقيه في العالم العربي من حيث الاحتكام الى صندوق الاقتراع والتاكيد على الديمقراطيه وسياده القانون. وتمنت التزام الرئيس الجديد بفترته المحدده باربع سنوات وعدم السعي الى تغيير القانون وان يعود للشعب المصري مره اخرى ان اراد تجديد ولايته لرئاسه ثانيه. وقالت د. المالكي ان المهام التي تنتظر الرئيس الجديد خاصه على الصعيد الداخلي شاقه وصعبه فهو ورث تركه ثقيله وهما ثقيلا والمهمه الملحه امامه هي اعاده ترتيب البيت الداخلي.

واعتبر عضو المجلس البلدي السابق ابراهيم ال ابراهيم ان فوز الدكتور محمد مرسي بالرئاسه اثبت نجاح مصر في اجتياز المرحله الصعبه التي مرت بها ولا تزال تمر بها وجنبها الوقوع في مزالق جديده. وراى ان المرحله المقبله بعد فوز المرشح الاسلامي الدكتور محمد مرسي لن تكون سهله، لكنه رجح قدره الرئيس الجديد على تجاوزها وبناء الانموذج الذي نامل ان يمتد تاثيره الى جميع الدول العربيه. واضاف ان نجاح الثوره في تحقيق اهدافها في مصر سينعكس ايجاب ا على جميع الدول العربيه وعلى البلدان العربيه التي شهدت ثورات الربيع العربي معربا عن الامل في ان تتغير الامور في العالم العربي نحو الافضل.

وراى الكاتب الصحفي عيسى ال اسحاق ان فوز الدكتور محمد مرسي بالرئاسه في مصر جنب الشعب المصري وقوع حرب اهليه في البلاد. وقال ال اسحاق نحن نامل ان ينجح الرئيس الجديد في انتشال مصر من الازمات التي تمر بها، وان يفي بوعوده في تشكيل حكومه وحده وطنيه تضم كافه اطياف المجتمع المصري. معتبرا ان ذلك هو التحدي الاساس امامه الان.واضاف ان جماعه الاخوان المسلمين طرحت خلال العقود الماضيه شعار »الاسلام هو الحل« وهم وصلوا الى سده الحكم في مصر الان وهي مطالبه بتحويل هذا الشعار الى برامج اجتماعيه واقتصاديه وسياسيه ليرى الشعب المصري نتائجه على ارض الواقع.ودعا ال اسحاق الرئيس المصري الجديد الى التركيز على القضايا الداخليه المصريه. مؤكدا ان قوه مصر وقوه الشعب المصري ستنعكس ايجابا بالضروره على الوضع العربي باكمله.

واعتبر الدكتور ربيعه الكواري ان تجربه الانتخابات الرئاسيه في مصر كانت ناجحه بدليل فوز مرشح الاخوان المسلمين. وقال الدكتور ال كواري ان التغيير والتجديد مطلوب وان على الرئيس الجديد ان يسعى في هذه المرحله الى تحقيق مبدا العداله الاجتماعيه بين المواطنين والقضاء على مشاكل الفقر والبطاله واعاده الحياه للاقتصاد المصري وان ينفذ البرنامج الذي انتخبه الشعب المصري على اساسه.

من جهته دعا الدكتور عيسى مطر الاستشاري في مؤسسه حمد الطبيه الرئيس الجديد الى ان يكون ملتصقا بهموم ومشاكل الناس وان ينزل الى الشارع ليسمع مطالبهم وقضاياهم. معبرا عن الامل ان يكون في نجاح الرئيس الجديد خير لمصر وللامه العربيه والاسلاميه. فوز مرسي يؤسس لبناء مصر الجديده دعوا الرئيس المصري لتحقيق الامن والاستقرار.

Template Summaries (Arabic Example)

{العنوان}: تحقيق الامن والاستقرار واعاده بناء مصر من جديد على اساس المساواه والعداله الاجتماعيه

{تاريخ النشر}: ٢٥/آغسطس/٢٠١٢

{الكاتب}: أنور الخطيب

{ الأشخاص المشار اليهم}: محمد مرسي, عبد الحميد الانصاري, موزه المالكي, ابراهيم ال ابراهيم, عيسى ال اسحاق, ربيعه الكواري

{ المؤسسات المشار إليها}: رئيس مصر, جامعه قطر, حزب الحريه والعداله , جماعه الاخوان ا لمسلمين, مؤسسه حمد الطبيه

{التصنيف العام}: السياسية

{الكلمات في الموضوع الرئيسي}: حكم, محمد , مرسي , الاخوان, المسلمين, مصر, الرئيس، والاستقرار ، السياسية، رئيس

Template Summaries (English Example)

{Title}: Achieve security and stability and the rebuilding of new Egypt on the basis of equality and social justice

{Publication Date}: 25 August 2012

{Writer}: Anwar Al-Khateeb

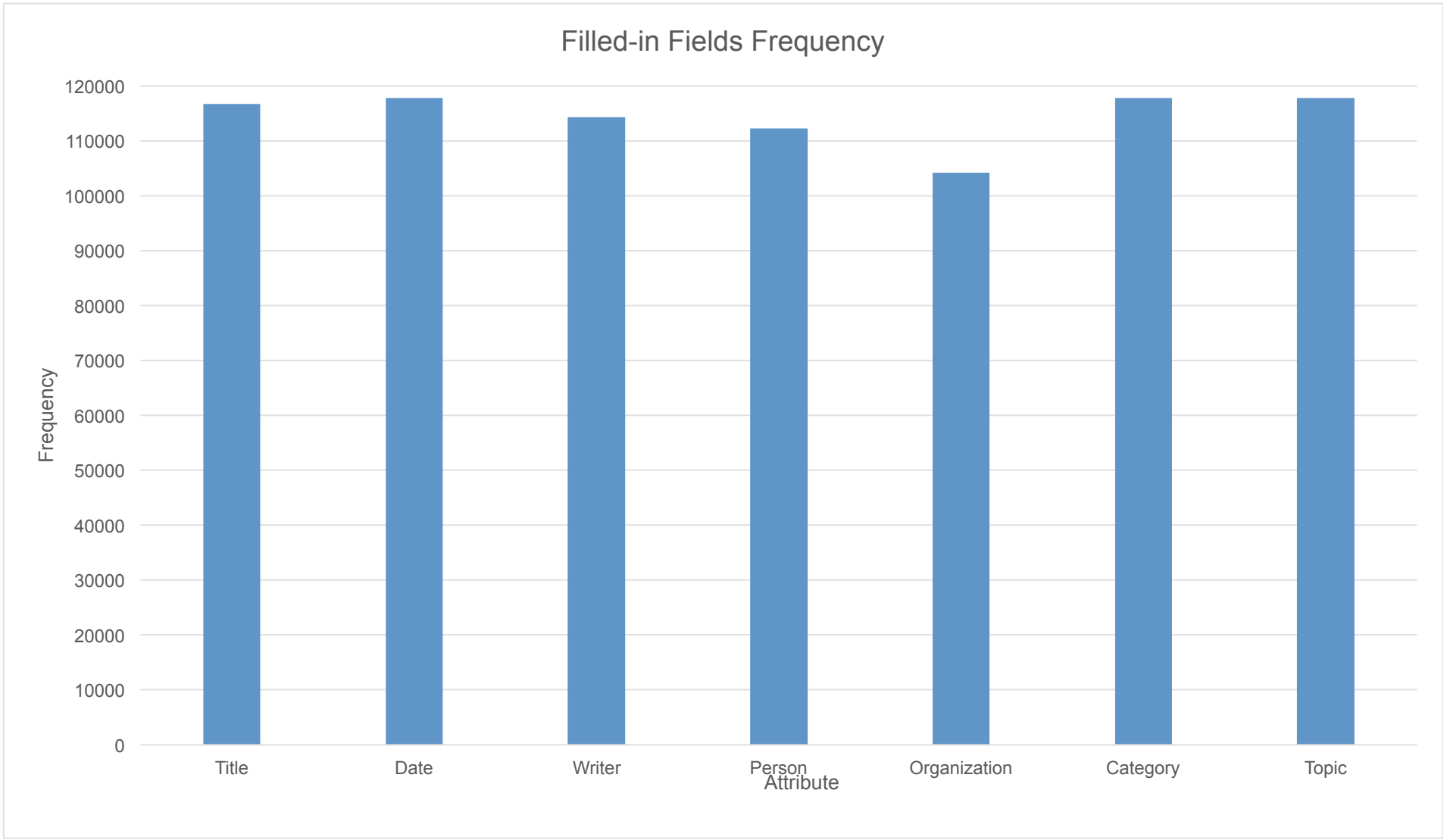
{People Mentioned}: Mohammed Mursi, Abdul Hamid Ansari, Moza al-Maliki, Ibrahim Al-Ibrahim, Isa Al Isaac, Rabia Al-Kuwari

{Organizations Mentioned}: Qatar University, Freedom and Justice Party, The Muslim Brotherhood, Hamad Medical Corporation

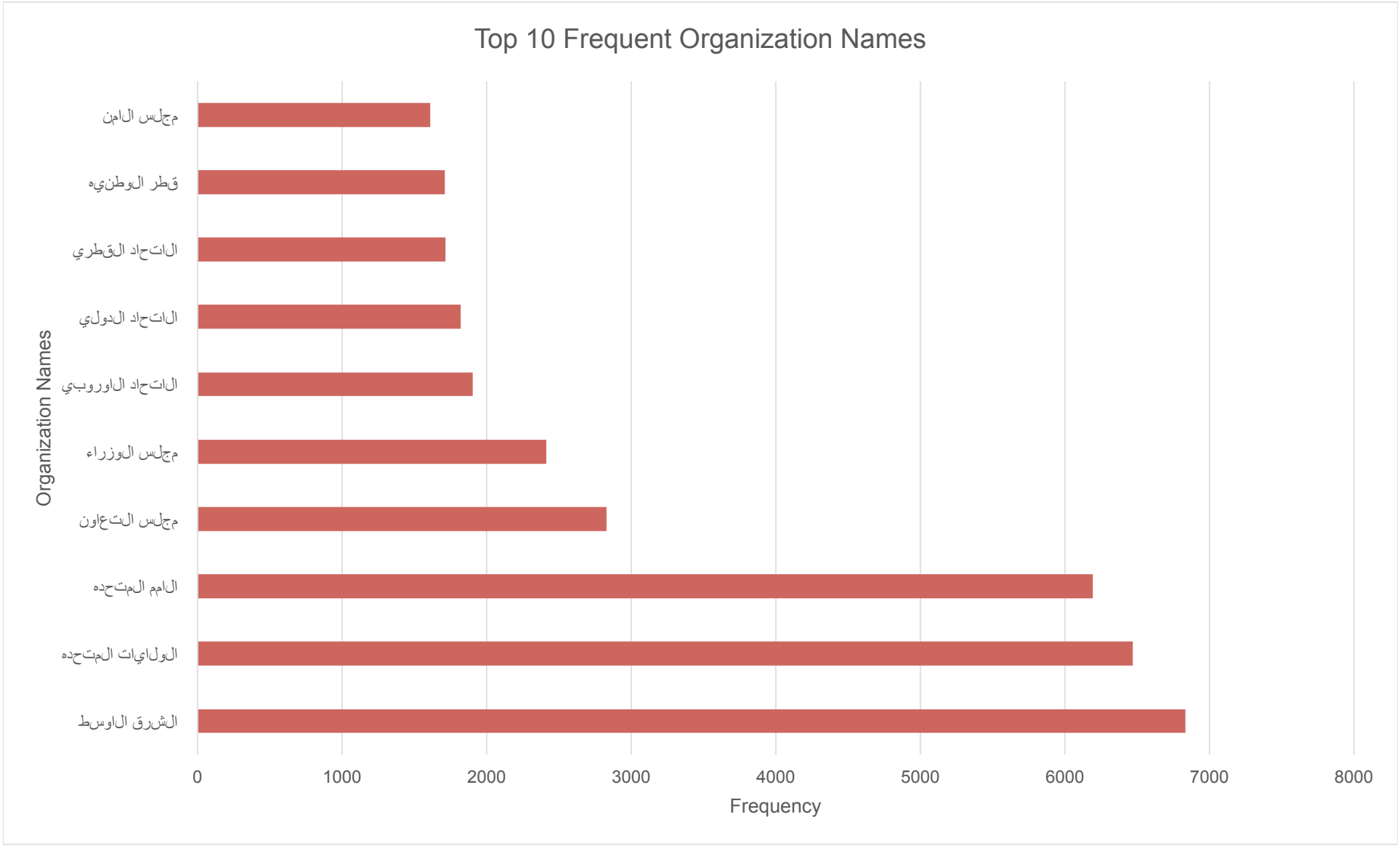
{General Topical Category}: Politics

{Words in Main Topic}: Leader, Muhammad, Morsi, Brotherhood, Muslims, Egypt, President, stability, political, Chairman

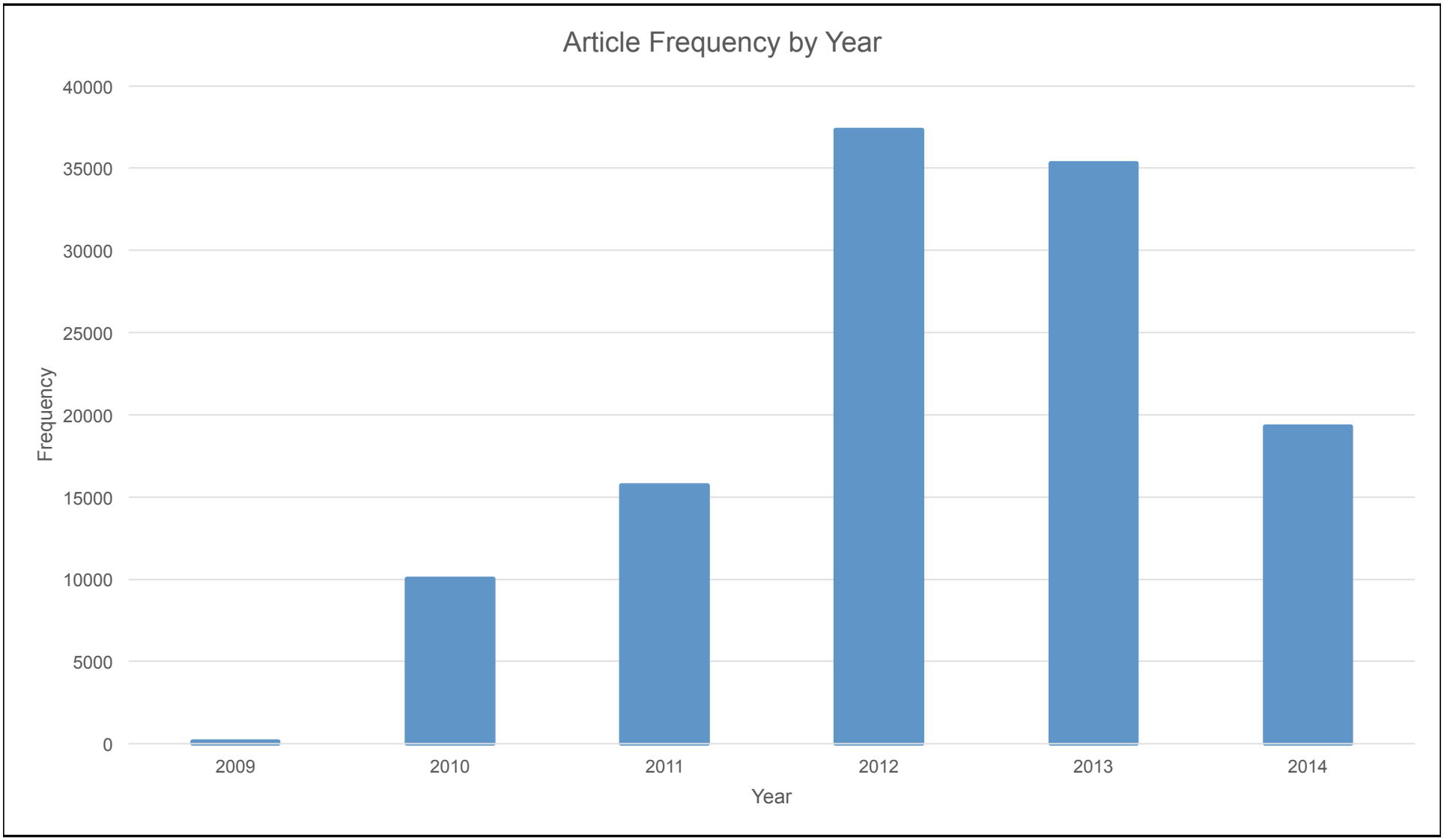
Sample Results Statistics



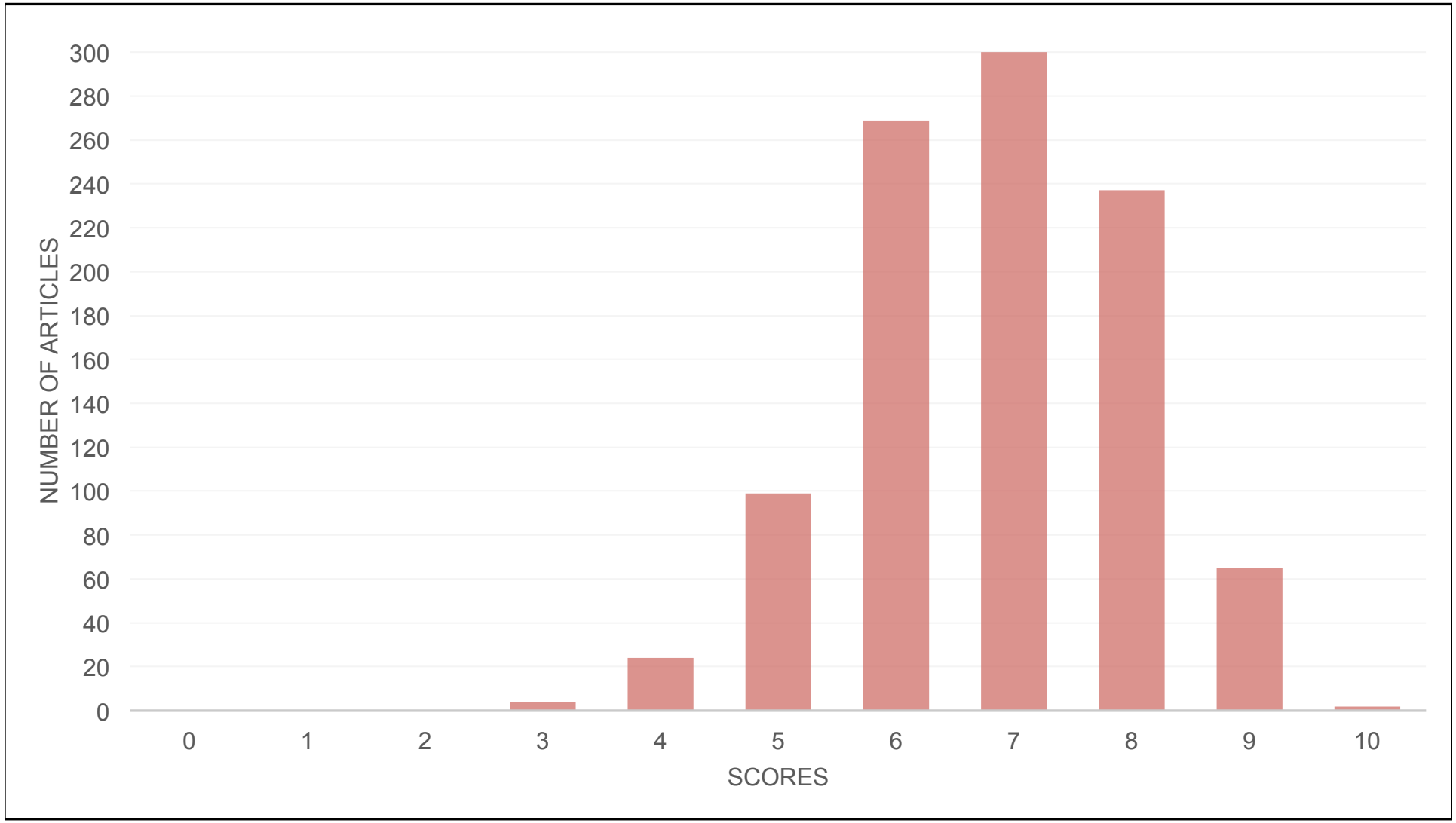
Sample Results Statistics



Sample Results Statistics



Template Summaries Evaluation



VirginiaTech

Template Summaries Evaluation

Rate Value	0	1	2	3	4	5	6	7	8	9	10
Number of Articles	0	0	0	4	24	99	269	300	237	65	2

Conclusions

Publications

Acknowledgments

Publications

- JASIST- Accepted for publication
- JCDL- Published
- ICSIC- Accepted for publication
- Final Results (Chapter 3)- Submission to a journal planned

Acknowledgement

- I would like to acknowledge
  - –My committee's help and suggestions with this work.
  - –My advisor Dr. Edward Fox for all his support and advice over my years at Virginia Tech.
  - –Mr. Philip Young, a scholarly communication librarian, for his help as a librarian expert
  - –Arabic native speakers, who served as experienced volunteers and helped with this work
  - –The group of graduate students who help build the baseline corpus and evaluate our results
  - –QNRF for their support. This research was made possible by NPRP grant # 4-029-1-007 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the author

Thank You!

Building a Standardized Categorization System for Arabic Newspapers

- Based on the ideas of
  - Topic **coverage**
  - **Common** categories
- We used
  - Taxonomy **graphical** mapping to build the taxonomy
- With the aim to:
  - Enhance Arabic news article **classification**
  - Improve online newspaper **browsing**

The Standardized Categorization System

- An ontology librarian **expert** and 5 native Arabic speaking **volunteers** have helped evaluate versions of our categorization system.
- I created the taxonomies for the 5 Qatari news papers and many other versions of our general and standardized taxonomy.
  - –Volunteers confirmed that each category was indeed representative of topics in news articles.
  - –The librarian expert
    - •Validated and approved its coverage
    - •Validated its cross-referencing against the IPTC system.

Arabic News Articles Text Classification

- The goal of a **root**-based stemmer is to extract the very basic form for any given word.
- The goal of a **light** stemmer is to find the canonical form of an Arabic word by removing prefixes and/or suffixes.

الاقتصادي = قصد

Word	Root
الاقتصادي The economic	قصد Meant
مقاصد Purposes	قصد Meant

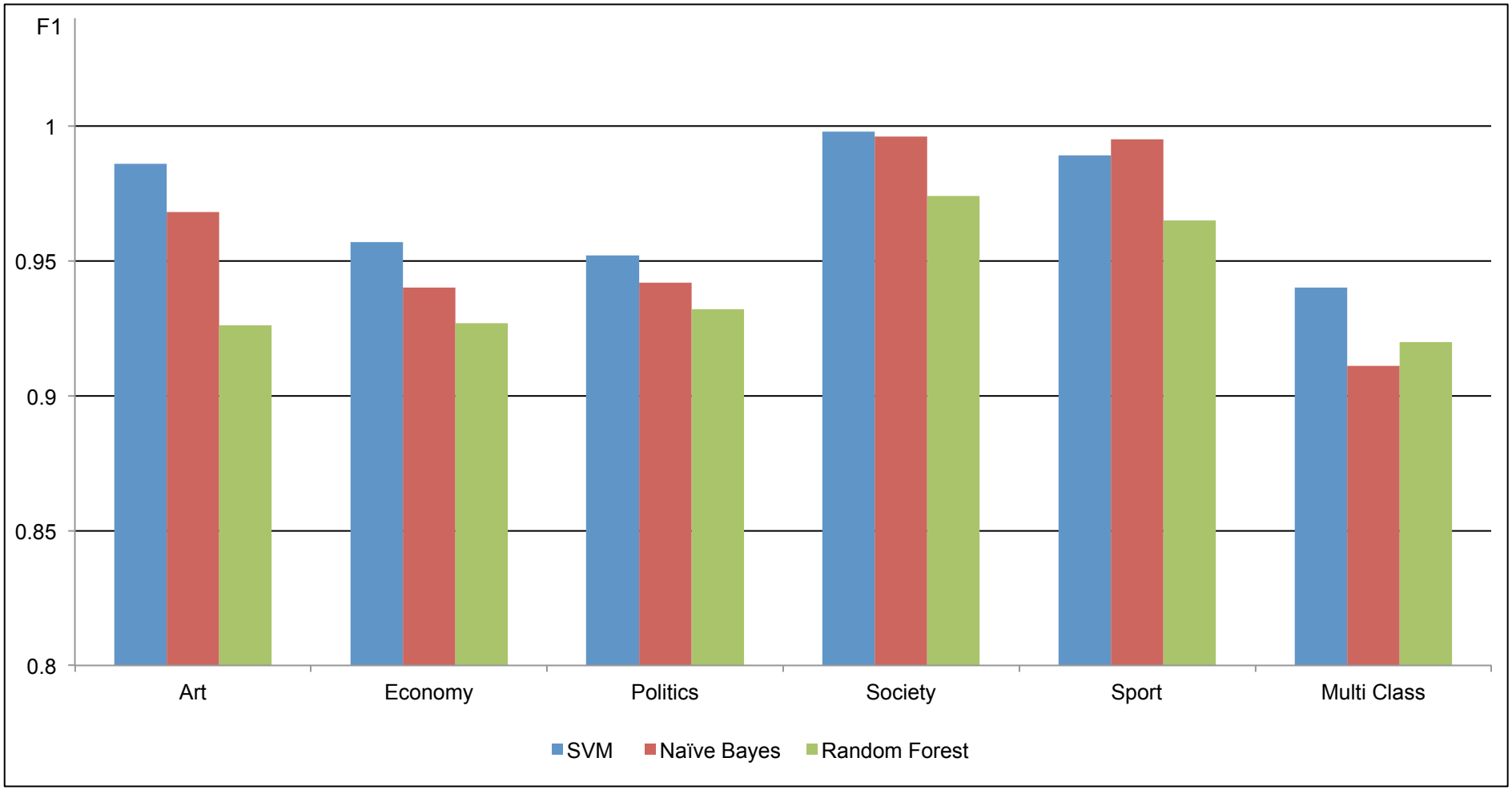
Word	Light Stemmer Result
اقتصادي Economic	اقتصاد Economy
والاقتصاد And the economy	اقتصاد Economy

Arabic Text Classification

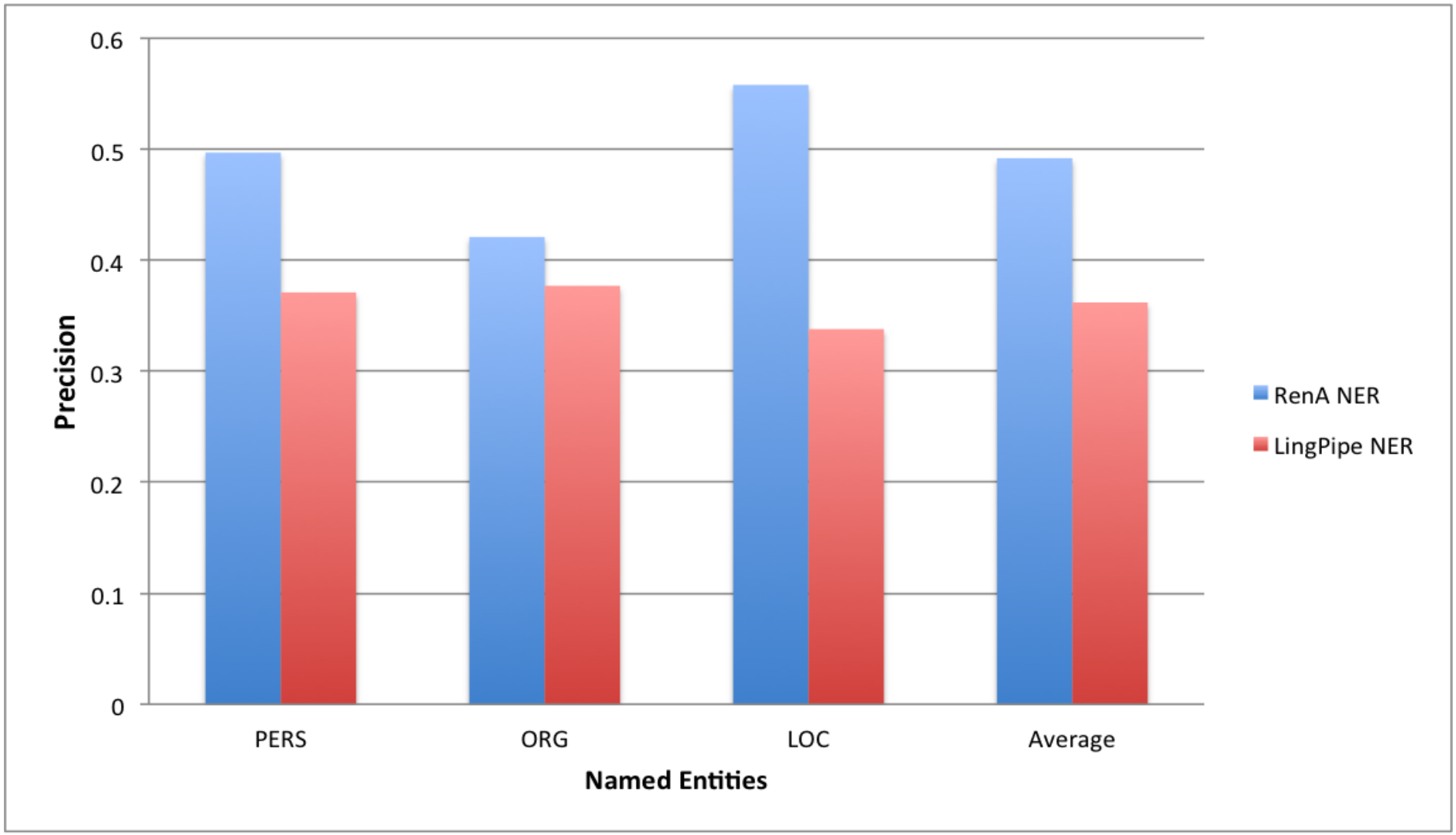
- We used the **SVM, NB**, and **RF** classifiers to
  - Judge the performance of the P-Stemmer
  - Compare it with the other listed approaches
- 10-fold cross-validation was used to train and test **36 classifiers**
  - 21 classifiers for multi-class (3 classifiers with 7 different data sets)
  - 15 classifiers for binary (3 classifiers with 5 different categories)

Arabic Text Classification

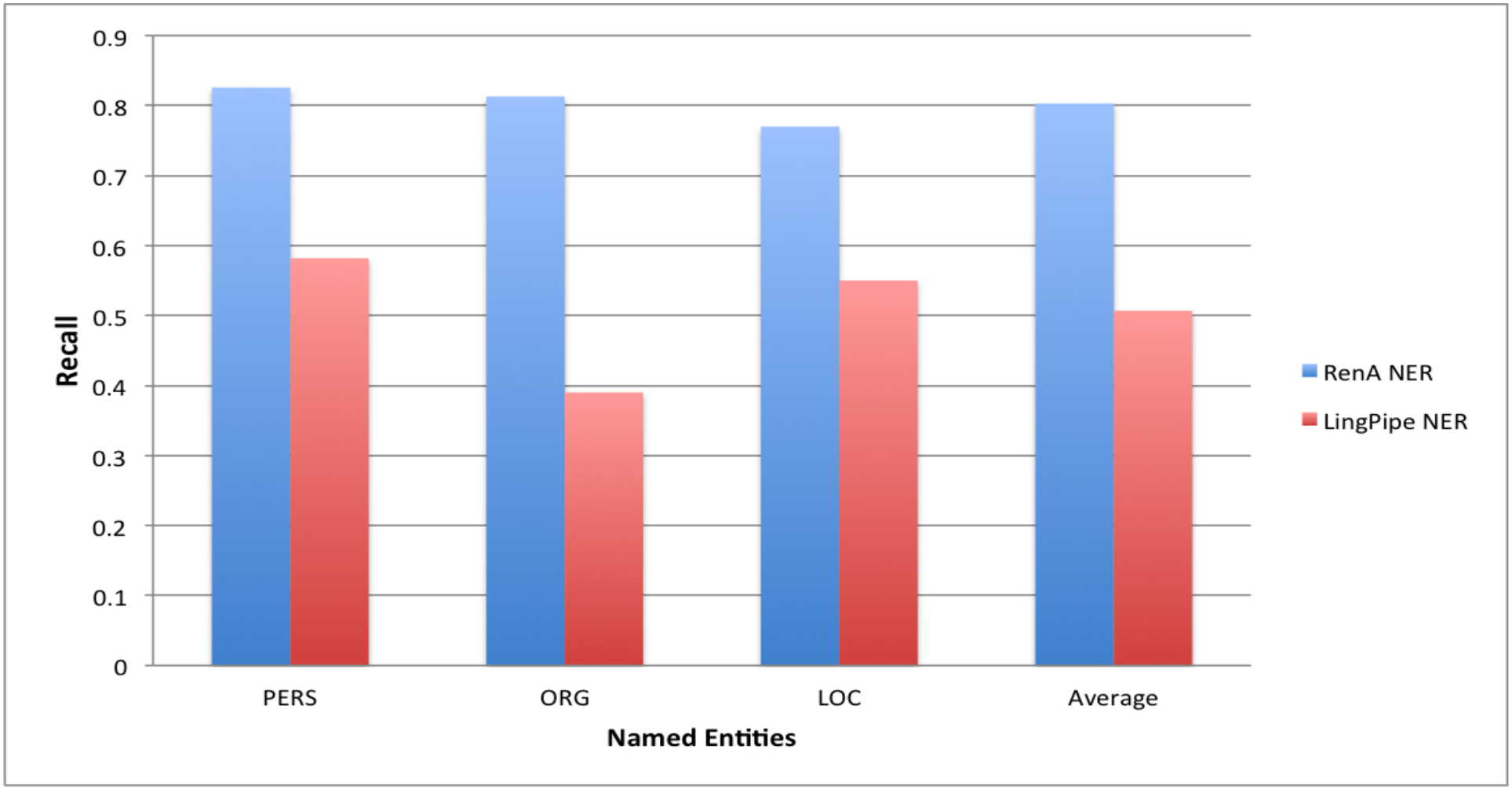
F1-Measure Values for the Three Classification Techniques with Respect to the Five Categories Training Sets



RenA Precision Results

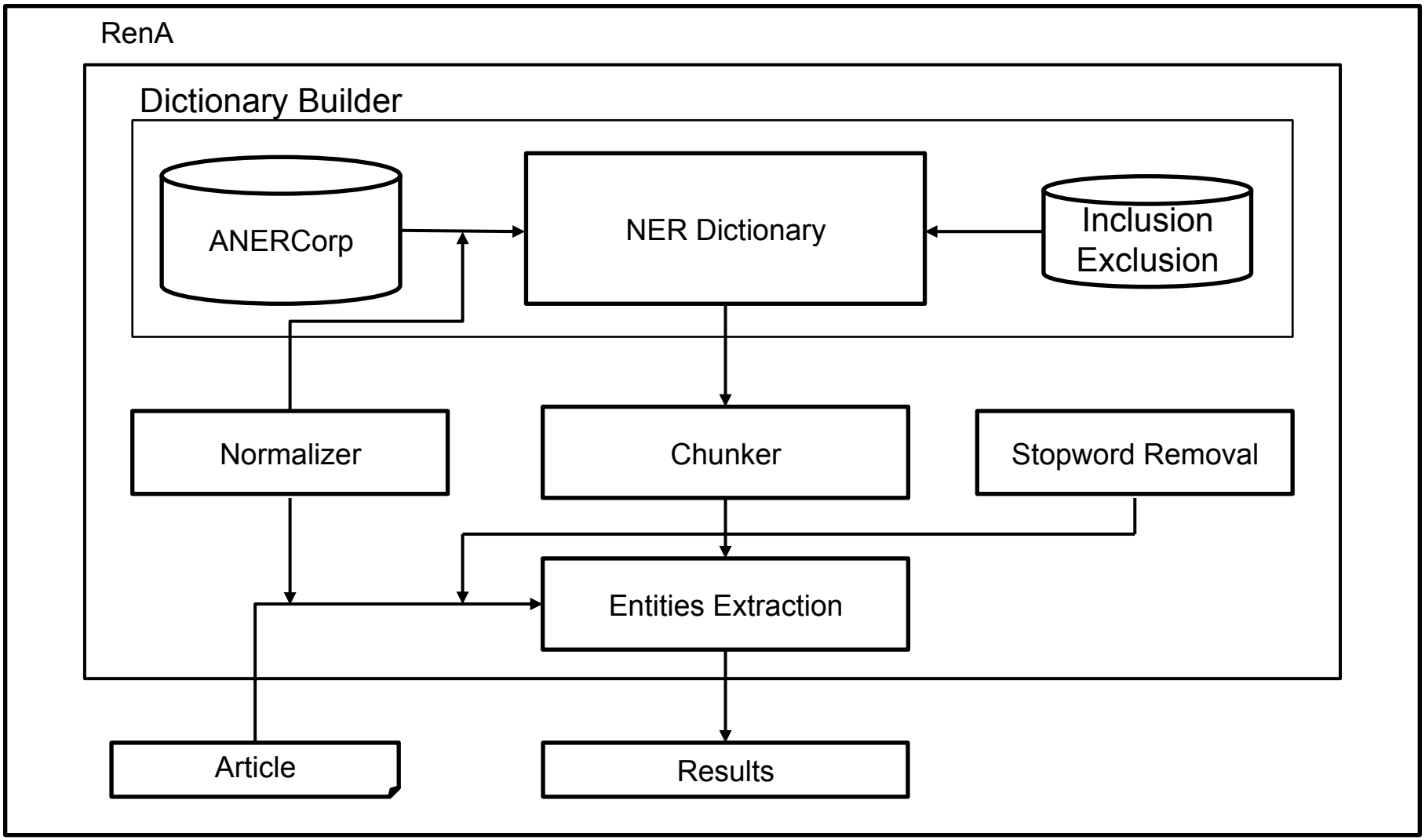


RenA Recall Results

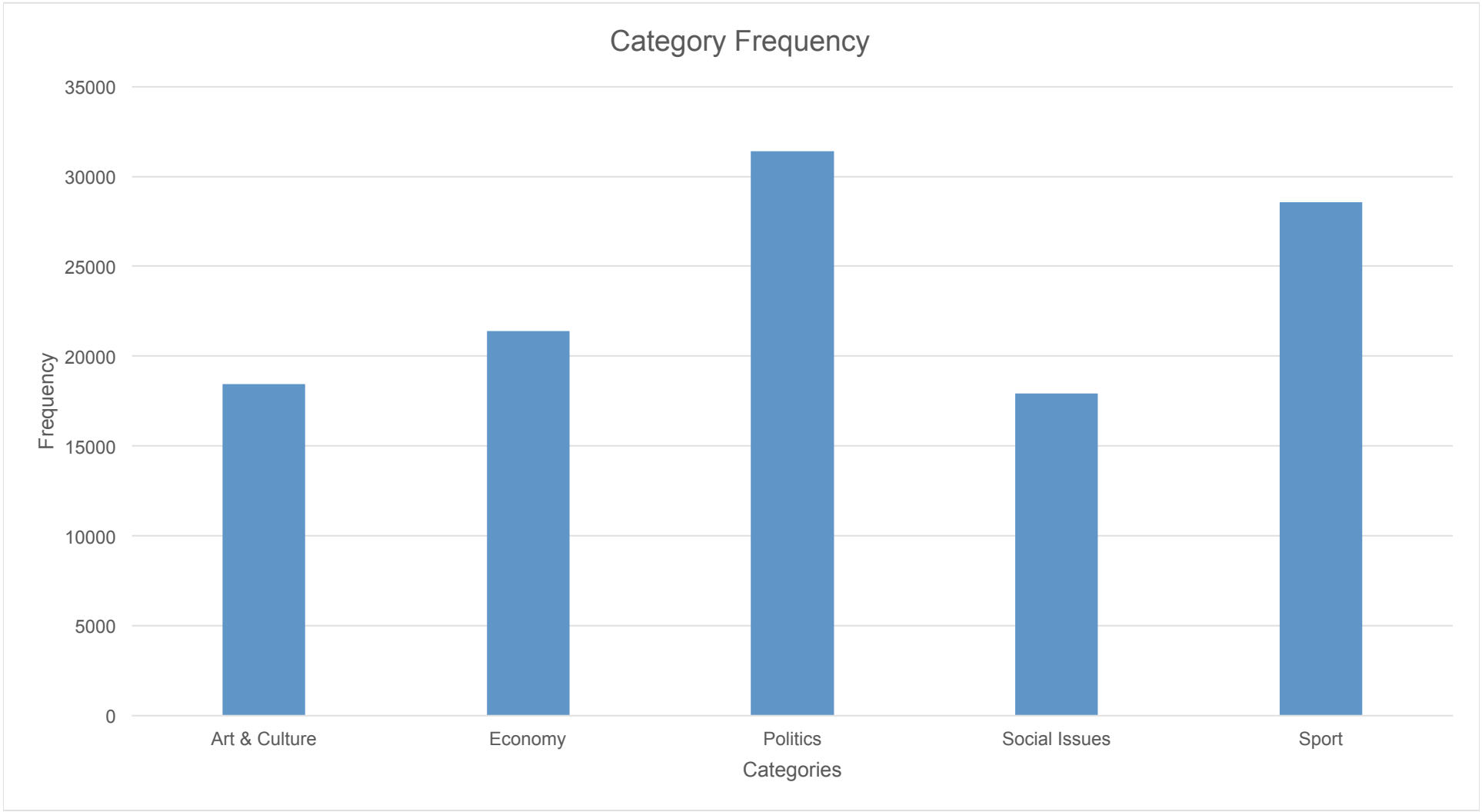


VirginiaTech

RenA


RenA
Dictionary Builder
ANERCorp
NER Dictionary
Inclusion
Exclusion
Normalizer
Chunker
Stopword Removal
Entities Extraction
Article
Results

Category Attribute Frequency



Frequency and Percentage of Missing Values for the Summary Attribute

Attributes	Count	Percentage
Title	1042	0.9%
Date	0	0.0%
Writer	3420	2.9%
Person	5462	4.6%
Organization	13553	11.4%
Category	0	0.0%
Topic	0	0.0%