Pattern Analysis of News Media Content

Ilias Flaounas

PhD Thesis

A dissertation submitted to the University of Bristol
in accordance with the requirements for award of the degree
of Computer Science in the Faculty of Engineering

October 25, 2011

Dedicated to my grandfather
Emmanouel Papoutsis who first
sparked my interest in science.

Preamble

Abstract

Our research is situated in the emerging field of Computational Social Sciences and more precisely on the intersection of Media Studies and Artificial Intelligence. News content analysis, as performed by media scholars, is based on the analysis of small datasets of articles that appeared in few outlets and during limited time periods. They are restricted by the manual methods of data analysis, known as 'coding'. Nowadays, the content of most news sources is available online. This fact, in conjunction with recent advances in machine learning, natural language processing and knowledge discovery in data, has enabled the automation of many aspects of social science research. Furthermore it made possible the answering of some questions about news media for the very first time due to the plethora of data that enabled a 'data-driven' approach to research.

In this research we undertake a large-scale textual content analysis of mainstream media using automated techniques. First, we developed the required infrastructure for the gathering and processing of textual media content from hundreds of news outlets, in multiple different languages and for extended periods of time. These data were processed in order to simulate the 'coding' that is performed by social scientists. Key findings include the comparison of topic selection bias of news outlets after automated categorisation of their content; the measurement of linguistic subjectivity and readability among different topics and among different outlets; the prediction of which articles have better chances of becoming popular before they are published. The network of "outlets that discuss the same stories" was created and analysed for the first time. We showed its stability in time and the predictability of its structure. Finally, we expanded to study the relationships among countries as these are reflected in their media content. We measured how factors such as geographic proximity, economic and cultural relations affect the selection of stories that are covered by media in each country.

Acknowledgements

The present research study would not have been completed without the help of many generous and inspiring people. At this point I want to thank and recognise for their valuable contributions the following individuals. I would like to express my deepest appreciation to my supervisor, Prof. Nello Cristianini not only for giving me valuable feedback, guidance, support and motivation throughout my academic endeavours, but also for setting the highest of standards for my academic work. Without question, Nello may well be, and has been, an inspiration throughout my study. However, at this point I would like to acknowledge that he has been a lot more than my supervisor during my studies.

The acknowledgements will be incomplete if I don't particularly thank for their valuable contributions the colleagues from the Patterns Analysis & Intelligent Systems research group which is part of the Intelligent Systems Laboratory at the University of Bristol. I am grateful to Dr. Tijl De Bie for his support and advice on data analysis; Dr. Marco Turchi for the implementation of the machine translation module that enabled the multilingual aspect of the analysis in my work; Omar Ali for creating and maintaining the spiders and the data collection pipeline that gathered the data used in this research; Elena Hensinger for our joint work on predicting popularity of articles; Nick Fyson for the collection of data used in measuring relations of countries as reflected in their media content; Tristan Snowsill for our work on demos; and Dr. Florent Nicart for the development of the first version of the API used in NOAM system. Thank you all for giving me the opportunity to collaborate with you and together to establish and make our work known to the academic community. I would also like to thank Dr. Philip Naylor for his support on the group's computer hardware infrastructure.

An important role for the accomplishment of this research study played Prof. Justin Lewis and Nick Mosdell from the School of Journalism, Media and Cultural Studies of Cardiff University for enabling me to acquire valuable insights regarding social scientists' perspective in relation to the medias study.

I would like to thank Tristan Snowsill, Omar Ali, Eirini Spyropoulou and especially Aikaterini Kokkinaki, for proofreading and providing valuable feedback during the writing of the thesis.

I want to acknowledge Alexander S. Onassis Public Benefit Foundation for supporting financially my PhD studies.

Last but not least I want to thank my parents, Nikolaos Flaounas and Marina Flaouna, for always being supportive in this long journey of mine.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE:.....

Publications

Most of the results of my research have been presented in international conferences (with full peer review) or published in scientific journals. The main text of the thesis contains references to the corresponding papers. It follows the full list of all papers related to this thesis and my contribution to each of them (P: wrote the paper, E: Performed Experiments, M: Contributed Analysis Methods, D: Contributed news content data and software for data acquisition):

1. **I. Flaounas**, O. Ali, M. Turchi, T. Snowsill, F. Nicart, T. De Bie and N. Cristianini: "NOAM: News Outlets Analysis and Monitoring System", Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, pp. 1275–1278, Athens, Greece, 2011.
   Contributions: P, M(HTML Scraper, Feed Finder, Topic Taggers, Clustering, Sentiment, Readability, System Watch), D — See Sect. 2.4

2. E. Hensinger, **I. Flaounas** and N. Cristianini: "Learning Reader's Preferences with Ranking SVM", International Conference on Adaptive and Natural Computing Algorithms, LNCS, Vol. 6594(2), Springer, pp. 322–331, Ljubljana, Slovenia, 2011.
   Contributions: M(Tagging), D — See Sect. 3.2.3

3. **I. Flaounas**, M. Turchi, O. Ali, N. Fyson, T. De Bie, N. Mosdell, J. Lewis and N. Cristianini: "The Structure of the EU Mediasphere", PLoS ONE, Vol. 5(12), pp. e14243, 2010.
   Contributions: P, E, M, D — See Chap. 6

4. **I. Flaounas**, N. Fyson and N. Cristianini: "Predicting Relations in News-Media Content among EU Countries", 2nd International Workshop on Cognitive Information Processing, IEEE, pp. 269–274, Elba, Italy, 2010.
   Contributions: P, E, M, D — See Chap. 5

5. O. Ali, **I. Flaounas**, T. De Bie, N. Mosdell, J. Lewis and N. Cristianini, "Automating News Content Analysis: An Application to Gender Bias and Readability", Workshop on Applications of Pattern Analysis (WAPA), JMLR Workshop and Conference Proceedings, Vol.11, pp. 36–43, Windsor, UK, 2010.
   Contributions: M/E(Readability), D — See Chap. 4

6. T. Snowsill, **I. Flaounas**, T. De Bie and N. Cristianini: "Detecting events in a million New York Times articles", Machine Learning and

Knowledge Discovery in Databases (ECML PKDD), Springer, LNCS, Vol. 6323(3), pp. 615–618, Barcelona, Spain, 2010.
Contributions: D (parsed data) — See Sect. 3.2.4

7. E. Hensinger, **I. Flaounas** and N. Cristianini: "Learning the Preferences of News Readers with SVM and Lasso Ranking", Artificial Intelligence Applications and Innovations, Springer, IFIP Advances in Information and Communication Technology, Vol. 339, pp. 179–186, Larnaca, Cyprus, 2010.
Contributions: P, E (SVM ranking), D — See Sect. 3.2.3

8. **I. Flaounas**, M. Turchi, T. De Bie and N. Cristianini: "Inference and Validation of Networks", Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Springer, LNCS, Vol. 5782(1), pp. 344–358, Bled, Slovenia, 2009.
Contributions: P, E, M (Inference/Validation), D — See Chap. 5

9. M. Turchi, **I. Flaounas**, O. Ali, T De Bie, T. Snowsill and N. Cristianini: "Found in Translation", Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Springer, LNCS, Vol. 5782(2), pp. 746–749, Bled, Slovenia, 2009.
Contributions: M (Tagging), D — See Sect. 3.3.1

10. **I. Flaounas**, M. Turchi and N. Cristianini: "Detecting Macro-Patterns in the European Mediasphere", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 527–530, Milano, Italy, 2009.
Contributions: P, E, M (Networks/Tagging), D — See Sect. 3.2.1

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Media System

The media system directly influences our society. It has been argued that people's view of the world is influenced more by the media than by personal experience [59, 117]. Arguably, the global media system has an important role in politics, economy and democracy [89, 34]. The public opinion and awareness are influenced by the way news are reported in media as the agenda-setting theory suggests [129]. Thus, news outlets need to be balanced in their reports and provide equal coverage to each and every opinion [82]. Fairness and accuracy of covered topics can also be considered equally important qualities. An additional responsibility of media in our society is to protect the public interest [59].

The media system is comprised of a diverse range of news outlets including newspapers, broadcast media such as TV and radio stations, magazines, news agencies, press offices of governments and corporations, *etc*. The set of all these news sources is also referred to as the mediasphere. The system operates in a global scale with information generated close to the source of events and then transmitted across the mediasphere, from one media to another, based on the choices of news editors.

In practice the media system operation is not ideal. As suggested by Herman and Chomsky in their celebrated propaganda model:

> "The societal purpose of media is to inculcate and defend the economic, social, and political agenda of privileged groups that dominate the domestic society and the state." [89]

The authors elaborate on this attitude by proposing a series of 'filters' which explain how media choose and shape the news. To give an example, the advertising filter suggests that media will not report negative news against corporations which provide their primary source of income. Another issue is the media's practice to have a constant coverage of some topics while systematically marginalising or completely ignoring others [130].

1.2 Media Analysis

Traditionally media analysis has been the domain of research of social scientists, particularly journalism scholars. From their perspective media analysis is viewed as:

> "The critical analysis of the various processes involved in gathering, evaluating, interpreting, researching, writing, editing, reporting and presenting information and comment on a wide range of subjects, that are disseminated via an expansive range of mass media to diverse audiences resident in local, regional, national and international settings." [71]

In their research they deploy various methodologies ranging from content analysis, conversation analysis, discourse analysis, agenda-setting to cultivation analysis [71, 118]. Their work is based on manual analysis and annotation of small sets of news articles, a process known as 'coding'. Although some commercial databases and dedicated software packages have been used to produce and analyse larger sample sets, journalism scholars still rely on time consuming manual processes of reading, coding and analysis of news. For example, two recently published studies in the *Journal of Communication* – currently one of the top journals in media studies – illustrate the order of magnitude of the number of outlets and articles they analyse. The first

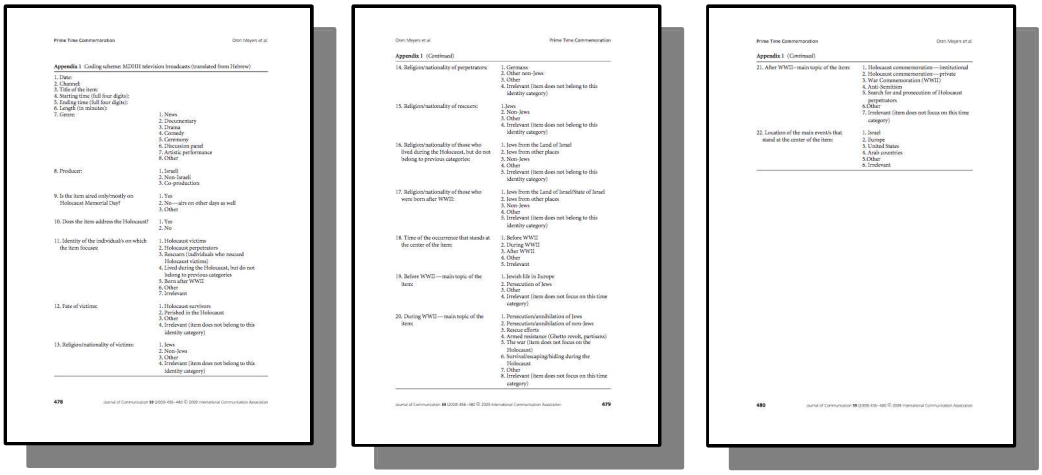


Figure 1.1: Examples of coding sheets that were used in [135].

study researches the way stories are reported in two broadcast media TV shows about two recent wars:

> "A total of *529* stories from NBC Nightly News and *322* stories aired on Special Report about Iraq, and *64* and *47*, respectively, about Afghanistan were analysed *by two coders*." [4]

The second study concerns the way Israel's Memorial Day for the Holocaust and the Heroism (MDHH) is reported in one main TV station:

> "Our corpus of data consisted of Channel 2s broadcasts on the eve of MDHH between 7:30 p.m. and midnight in the years 1994-2007 [...]. All *278* items aired on the *14* examined evenings were *coded*." [135]

From the studies mentioned, it becomes apparent that not only the limited amount of data is an issue but also the time-consuming process of manually annotating it. Figure 1.1 presents an example which illustrates the three pages questionnaire that coders completed for each of the news items that they were analysed in [135]. In total 22 questions were asked including, *e.g.*, "Does the item address the Holocaust?", "Location of the event/s that stand at the centre of the item", or the main topic/s addressed in the item before,

during and after the WWII. Coding can provide significant insight, when small numbers of outlets are analysed for a limited period of time and high accuracy is sought.

However, the opportunity now exists to monitor a vast number of outlets constantly and in an automated way. Nowadays, news from most mainstream media is freely and easily available in digital format via the web. This wide accessibility of news has reached the point of characterising news as 'ambient' [84]. The required automation of analysis can be realised by utilising Artificial Intelligence techniques and more precisely methodologies from the fields of Machine Learning [16, 178], Data Mining [119, 32] and Natural Language Processing [126]. These fields have now reached the level of accuracy where they can reliably be deployed to enable scientific analysis of global patterns in the content of the media system in a quantitative manner. Note that, even though the machine-analysis of a single article can still have some probability of error, the task of detecting patterns in large-scale corpora can lead to statistically significant results.

In this research we undertake a large-scale mainstream media textual content analysis using automated techniques:

- 'Large-scale' since, unlike social studies, we analyse typically hundreds of outlets in parallel, for extended periods of time, involving millions of news items.

- 'Mainstream' news media since we don't focus on alternative media, such as blogs or Twitter, but traditional news-media such as newspapers, magazines, broadcast media and so on.

- 'Textual' since we use only the textual information of news rather than other information, *e.g.*, images, videos or speech [112].

- 'Automated' in the sense that the analysis is performed by applying Artificial Intelligence techniques rather than using human coders as the standard approach of social scientists implies.

1.3 Related Work

The interaction of computer science and physical or biological sciences has been significantly successfully the last decades and has lead to revolutionary research in those fields, such as the CERN or the human genome projects. The interaction of computer science and social sciences has been much slower and currently it is considered as an emerging research field known as 'Computational Social Sciences' [111]. This delay is due to the complexity of the social interactions involved as well as the unavailability of digital data [189]. This interaction of social and computer sciences leads social scholars to step beyond the reach of traditional methods of analysis [35]. Advances from the fields of Data Mining and their interaction to real life allow now new types of questions to be asked and answered [140]. Recently social scientists started to pay attention to Machine Learning approaches (see, *e.g.*, [29]). Recent works that resulted from the field of Computational Social Sciences include: the large scale person-to-person interactions of people using RFID devices [31]; the study of friendships network structure and human interactions using mobile phone data [48, 143, 154]; the individuals human mobility patterns using mobile phone data [79]; human travelling patterns by analysing circulation of bank notes [19]; the discovery of patterns in email exchange [49]; the study of human interactions in online games and environments [12, 175]; the prediction of behaviour of various Techno-Social systems [187].

Our work is situated in the intersection of media studies, a subfield of social sciences, and several approaches of Artificial Intelligence. This is depicted in Fig. 1.2. The large scale analysis of media, in order to achieve interesting results, is a challenging task and requires the development and integration of diverse computer science methodologies ranging from Machine Learning, Data Mining and Natural Language Processing up to Data Management and Web Technologies. There are few systems that have been developed for the continuous media monitoring.

The 'Lydia' system [121, 14] is a multi-purpose system focusing mainly on US media. It has been used for detecting spatial and temporal distribution of named entities [134, 122] in the news; the comparison of entities

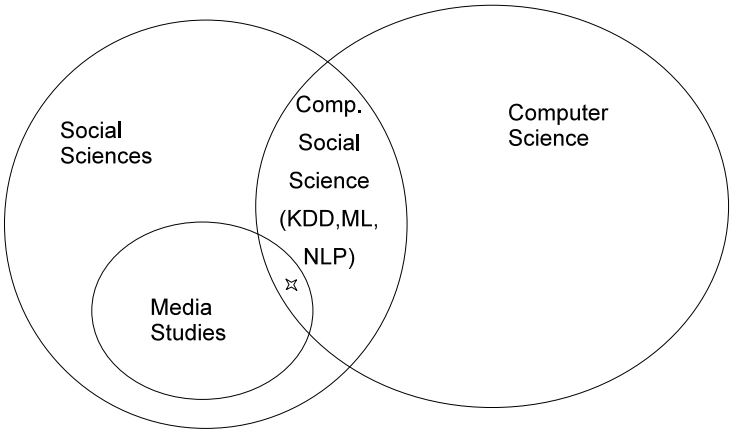


Figure 1.2: The area of our work is denoted by a star.

distribution in newspapers and in blogs [120]; the sentiment associated with entities as they appear in the news [13, 77]; the correlation of news reports about companies and their stock trading volumes and financial returns [196].

Another system is the Europe Media Monitor (EMM). It is supported by the European Commission [169] and it focuses mainly on the multilingual EU media. It has been used for multilingual event extraction [92]; the reconstruction of social networks of people [24]; the recognition and disambiguation of Named Entities such as places [21]; the detection and tracking of quotations in news [23]; the multilingual Named Entity recognition [22]; the continuous presentation of news in an integrated environment.

Other systems such as 'Newsblaster [132] and 'NewsInEssence' [153] both focus on generating multi-document summarisation of the daily news, although they monitor a limited number of news outlets. In the list of related work we could also add commercial systems such as Google News and Yahoo! news. These systems have the purpose of aggregating and presenting the news from multiple sources in an integrated and user friendly way.

We have a different goal to those media monitoring systems, and that is to detect patterns that can provide insight of the structure of the media system itself. This means we are primarily interested in studying the media system *per se* rather than presenting news to some audience. Also it is worth mentioning that, in general, we perform multilingual text analysis of outlets

from many countries globally and we don't restrict the monitoring to specific geographical areas.

Related to our project can be considered works that analyse news in blogs and the blogsphere. Blogs, and especially their relation to traditional media, are also in focus of analysis by social scientists [8, 128, 188]. Examples of computer science based studies of blogs include: the comparison of blogs during U.S. Elections [1]; the modelling of news flow in the blogsphere [3, 2]; the analysis of the blogs network [113]; the detection of communities in blogs [108]; events detection from the blogsphere [148].

Finally we need to point out that datasets commonly used as benchmarks in text mining research, such as the Reuters [116] and The New York Times Corpus [157], are not fully relevant to this project since our work is concerned with combining information from multiple sources. Nevertheless, we used them to build reliable topic taggers (Sect. 3.1.1). Other commercial databases, like 'LexisNexis'[1], commonly used by social scientists, provide a very limited human oriented interface that can be used primarily to query for articles using specific keywords. Similarly, commercial software, such as 'Atlas.ti'[2], can be used to ease the coding process and the qualitative or quantitative analysis of news, but they do not offer any large scale automation capabilities.

1.4 Key Findings

In this section, we summarise the key findings of our research, that are not within reach of traditional approaches applied by journalism scholars:

- The detection of laws that dominate the media content like the power-law distribution of number of outlets covering a story, or the seven days cycle on the volume of published stories.

[1] LexisNexis: Commercial database of news content. Accessible at http://www.lexisnexis.com

[2] Atlas.ti: Commercial Software: http://www.atlasti.com/

- The comparison of countries based on the topics their media choose to cover in a constant media monitoring manner.
- The large-scale comparison of outlets based on their topic selection bias.
- The large-scale measurement of numerical properties of news, such as linguistic subjectivity and readability, among different outlets as well as between different topics. This leads to a ranking of outlets and topics based on these properties.
- The prediction of those articles which have better chances of becoming popular before they are published.
- The inference of the network of news outlets that cover the same stories. We validated the network by showing its stability in time and the predictability of its structure.
- The measurement of factors such as geographic proximity, economic and cultural relations which affect the selection of stories that are covered by media in each country.
- The measurement of factors that correlate to the deviation of media content of countries from the 'average' media content.

1.5 Thesis Outline

In Chapter 2 we introduce NOAM, our system for news mining, that was developed for the gathering and analysis of news media content. Chapters 3 and 4 deal with approaches similar to those conducted by media scholars, such as coding and comparison of outlets. Chapters 5 and 6 analyse the media system in a large scale that goes beyond the reach of traditional methods.

More precisely the rest of this thesis is organised in the following chapters:

- **Chapter 2. News Mining** We introduce NOAM, the system that was developed for the gathering and annotation of news items. The

core of the system is an annotated corpus and an interface that allows media content to be queried at a semantic level. Several subsets of this corpus are used in this research to answer questions about the media system. In this chapter we also present basic techniques from the fields of Data Mining, Natural Language Processing, and Machine Learning that will be used throughout the thesis. Finally we show some first characteristics of the media system as they are captured in our corpus: the seven days periodic behaviour of the volume of published articles and stories, the power law governing the number of articles that form news stories, and the power law governing the number of news outlets that cover the same news stories.

- **Chapter 3. Machine Learning for News Analysis** In this chapter we present some classic Machine Learning methods that can be applied to the analysis of news media content. We utilise kernel methods for the automation of categorisation of articles based on their topic, a process often performed by social scientists manually. We also show that we can predict which articles are more appealing to a given audience and furthermore, we automatically generate lists of keywords that are expected to trigger the attention of readers. Finally we present some web based applications for demonstrating the ability of constantly monitoring the news media system.

- **Chapter 4. Quantifying Properties of News Items** In this chapter we measure numerical properties of news, namely readability and linguistic subjectivity. We compare US and UK newspapers based on these two writing style properties. We also compare general news topics in terms of their readability, linguistic subjectivity and popularity. This set of results illustrates how pattern analysis technology can be deployed to automate tasks similar to those performed by humans in the social sciences and the feasibility of large scale studies that would otherwise be impossible.

- **Chapter 5. Network Inference from Media Content** The theme

of this chapter is the modelling of the media system as a network of outlets. More precisely we introduce methods that can be applied for the reconstruction and validation of networks. We apply them for the network of the news outlets that cover the same stories and we visualise it for the first time.

- **Chapter 6. Patterns in the EU Mediasphere** In this chapter we extend our research to study relations among countries as they are reflected in their media content. We focus on a well defined subset of the leading media of the EU countries. While independently making a multitude of small editorial decisions, the media of the 27 countries shape the contents of the EU mediasphere in a way that reflects its deep geographic, economic and cultural relations. Detecting these subtle signals in a statistically rigorous way would be out of the reach for traditional methods. This analysis demonstrates the power of the available methods for significant automation of media content analysis.

- **Chapter 7. Conclusions** This chapter contains a discussion of the potential impact of the application of modern AI methods on media studies. We discuss the data-driven approach over the typical hypothesis driven scientific method. Finally we discuss some future avenues for further research.

Chapter 2

News Mining

We start this chapter by defining some basic concepts and terms such as the notion of news outlets, tags, feeds and news items. Next, we introduce 'NOAM', our own media monitoring system that gathers and analyses news items from the web[1]. We present the basic news mining, natural language processing and text processing techniques that were utilized in our system. Finally, we show some basic properties of the media system that are reflected in our corpus.

2.1 News Outlets

Informally a *news outlet*, or simply outlet, is a media company or organisation that provides news. Typical examples of news outlets include newspapers, magazines, broadcast media such as TV and Radio stations, blogs, newswires and so on. In this research we focus only on media that offer their content online, under the realistic assumption that nowadays the majority of influential news media will have an online presence. Figure 2.1 illustrates an example of an online article published in the webpage of BBC.

More formally we define a news outlet as an online source of news with a unique domain name. Under this definition, organisations with multiple

[1]The NOAM system is the product of extensive collaboration involving several people. While we present all of it here as part of the data collection and analysis infrastructure, we will emphasise and cover those modules that have been created as part of this thesis.

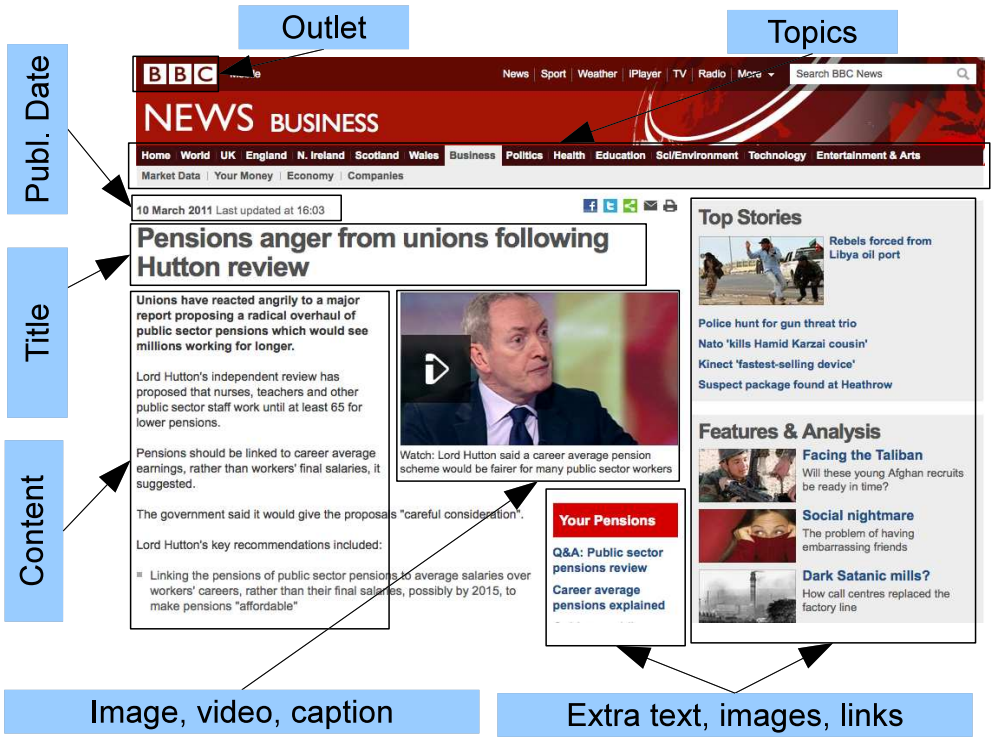


Figure 2.1: Example of an online article as published in BBC webpage. It is comprised by several parts which we highlight.

media under their control but with a single domain name are treated as one media. For example, BBC1, BBC2, BBC Radio, etc. are all treated as one outlet since they share the same domain name, *i.e.*, 'bbc.co.uk'. Blogs are an exception to this rule since they are treated as different outlets. This is because multiple Blogs can be hosted under the same domain name, *e.g.*, under domain name 'blogspot.com'.

2.1.1 Ranking of Outlets

In some parts of our research we will need to restrict statistical measurements only to a subset of news outlets to avoid introducing biases. This subset, usually in a per country basis, should contain the most influential media of each country. Thus we need to define a ranking of outlets based on their influence.

Objective rankings are based on opinion polls and circulation numbers of printed media. These measurements usually are not publicly available, they are restricted only to the media of a given country and type. Examples of organisations and companies that produce such data include the 'Audit Bureau of Circulations' for newspapers and magazines, the 'Broadcaster's Audience Research Board' for UK TV stations, and the 'Radio Joint Audience Research' for UK radio. Integration of these figures to generate an objective ranking of media is practically infeasible. How can differences in impact between broadcast and printed media be compared?

To overcome this problem, since we focus only on outlets that have online presence, we used as a measure of their impact the number of visitors to their websites. Exact numbers of visitors are not publicly available but estimations of websites' traffic are freely reported by companies such as Alexa.com (Alexa Traffic Rank index). This ranking is calculated using a combination of criteria and it estimates the traffic rank of each website. We collected the rankings of media websites of interest and used these web traffic rankings to rank the news outlets of each country. The same strategy has been successfully used before in literature, *e.g.*, in the spatial analysis of named entities in US newspapers [134].

2.1.2 Outlet Tags

The notion of tags is used widely in our research. A tag is typically a small piece of text that annotates an object (*e.g.*, a news item or outlet) and conveys some information about its carrier. We distinguish different, although related, 'tag-spaces' depending on the objects that the tags are associated with, *e.g.*, the space of tags for outlets, for news feeds or the space for news items. The relations between Outlets, News Feeds, Articles and their tags are illustrated in Fig. 2.2 and we will further elaborate on them in the next Sections.

A set of tags is associated with each outlet in our system. The typical set of tags that an outlet carries is the name of the country of origin of the outlet and its media type. We used the following media types: 'Newspaper';

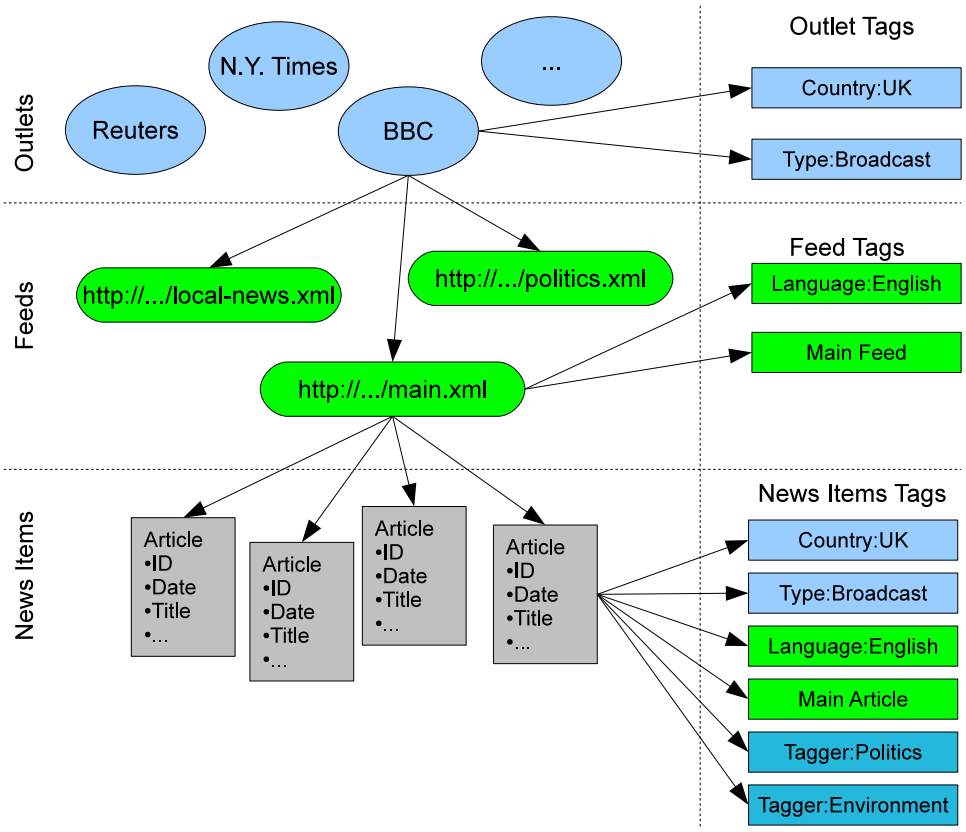


Figure 2.2: Example of the relations between Outlets, Feeds, News Items, as well as their corresponding tag-spaces.

'Magazine'; 'Broadcast media' (includes radio and TV stations); 'Newswire' (News agencies that provide news for other outlets); 'Blog' and 'Online media' (that is media like news portals that have an online-only presence and are not associated with mainstream media such as a newspaper).

2.2 News Feeds

News Feeds are XML-based technologies used to publish online frequently updated documents like blog entries and news articles. There are two major formats: the Really Simple Syndication (RSS) and Atom. Nowadays the majority of media outlets offer their content in one of those formats organised

Figure 2.3: Example of the main news feed published by BBC. It is a list comprised by several articles.

in one or more news feeds. Figure 2.3 illustrates as an example the first articles advertised in the main feed of BBC News. The corresponding XML code that we parse is shown in supplementary Fig. 8.1.

A web crawler infrastructure was developed in order to gather news feeds of outlets[2]. The crawling targets are based on a manually compiled list of feeds of interest. The discovery of new feeds can be automated as shown in the next section, but their manual annotation is still necessary. Further discussion on crawling of feeds can be found in [119].

2.2.1 Automatic Discovery of Feeds

Even the discovery of links to feeds of large commercial media sites is tedious work due to the large scale and the fact that we are dealing with websites in many different languages. The automatic discovery of news feeds to help this process is highly desirable. A crawler, namely 'Feed Finder' was developed that, using some starting points, can follow links and discover new feeds – this is a different crawler than the crawler that parses the feeds. Starting points are either specific media pages, or sites with links to multiple media pages. Heuristics have been applied in order to rank links and follow only the most promising ones. Such heuristics include the presence of terms such as

[2]The web crawler for the collection of news out of news feeds was developed and maintained by Omar Ali.

Table 2.1: Feed tags that are inherited from feeds to articles and the corresponding number of annotated articles. We do not show language, location and media type tags.

Tag Name	Articles	Tag Name	Articles
Acc. Disasters	4165	Most Popular	317940
Biology	51432	Museums	89
Business	1170577	NGO	60280
Chemistry	36486	Nuclear Weapons	327
Crime	49625	Op.-Ed.	335347
Democrat	378758	Physics	110607
Environment	66929	Politics	1683570
Events	7376	Religion	1049
Female Sports	7124	Republican	197148
Goverment	49230	Research Org.	118819
Health	420796	Space	53724
Hirrucanes	511	Sport	175912
Jobs	56276	Technology	484383
Life Science	51411	Terrorism	570
Local News	5083629	Think Tank	62265
Math	12650	Main Feed	24691444
Middle East	106	War And Conflict	4645
Most Emailed	41688	World News	4494125

'rss', 'feed', 'xml' in the link text or in the link URL. A standard depth-first algorithm visited all promising links up to a specific depth, while keeping a record of visited webpages, in order to increase efficiency. Currently over 223,360 feeds have been discovered using this approach. The speed of the discovery of new feeds can be further improved using Reinforcement Learning [172] as discussed in [155]. Supplementary Fig. 8.2 illustrates the web based front end of Feed Finder. User can add crawling jobs to a queue and navigate the discovered feeds.

2.2.2 Feed Tags

A set of tags is associated with each feed. Examples of such tags include: the language of the content of the feed, the tag about the outlet that carries

the feed and the country it belongs to. Tags such as country and outlet are inherited from the outlet to the feed. Language is a characteristic of the feed and not of the outlet, since some outlets have multilingual content and dedicated feeds for each language. The tags we use to annotate feeds and the number of corresponding articles are enlisted in Table 2.1.

Some publishers divide their content in a topic-based manner, and provide specialised feeds that carry the articles of each topic. We annotate such feeds with a corresponding topic tag, *e.g.*, 'Business', 'Politics', or 'Sports'.

A special tag, namely 'Top Stories' is used to indicate the main feed of outlets. This is the feed that is advertised in the homepage of a news media and it contains the main news items of the day, as they are selected by the outlet's editors. This feed is useful when a comparison among different outlets is required.

Other specialised tags include the 'Most popular' tag that annotates feeds that carry the most popular stories as they are formed by the readers' click choices. Such feeds are not present in all outlets. We use them in section 3.2.3 to analyse the popularity of news.

2.3 News Items

We use 'news item' as a general term that describes a news report that appeared in some outlet. It can have the form of a newspaper article, the transcription of a newscast in the TV program, the written work that appeared in a blog, *etc*.

News items can be seen as structured objects that contain a title, a description (which can be seen as a summary of the main article) and the full text of the article. Furthermore, each item has specific annotation such as publication date, the outlet it comes from, the URL that contains the full article, and also possibly a set of tags. The information associated with a news item is summarized in Table 2.2.

Let us give a specific example of an article from its advertisement in a news feed up to the point were it is collected and stored in our system. Initially the crawler visits the main news feed of BBC. That feed is illustrated in

Table 2.2: The components of a news-item.

Component	Description
Title	The title of the article, collected from the news feed.
Description	The description is a summary or the first few sentences of the article body and is collected from the news feed.
Content	The textual body of the article as gathered from the HTML scraper using the link that is provided in the news feed (See Sect. 2.3.1).
Date	The date that the article was first collected.
Link	The link to the main article that is advertised in the news feed.
Outlet	The outlet that published the article.
Inherited Tags	Tags that are passed to the item from the outlet tag-space and from the feed tag-space (See Sect. 2.3.2).
Content Based Tags	These are tags that are assigned in an automatic way from modules of NOAM (See Sect. 2.4).

Fig. 2.3 and the corresponding XML code in Fig. 8.1. System collects Title, Description and Link. The scraper follows the link in the XML file to the actual article page, illustrated in Fig. 2.1, and collects the main article body. The feed we crawled is annotated with the tags 'English' and 'Business', while the outlet with tags 'Broadcast' and 'UK'. These four tags are inherited to the news article. Next, the crawler visits the rest of the news feeds of BBC and discovers the same article in the specialised feed of BBC on Business news. The article is not stored for a second time, but the Business feed inherits to the afford collected news item the extra tag 'Businness'. All the information stored about the article is illustrated in Table 2.3.

Table 2.3: Example of a news item published in BBC.

Component	Value
Title	Unions' anger over pensions plan
Description	Unions react angrily to a report proposing a radical overhaul of public sector pensions which would see millions working for longer.
Content	Millions of workers in the public sector should work longer for lower pensions, a major report has said. [...]
Link	http://www.bbc.co.uk/news/business-12687489
Outlet	BBC
Inherited Tags	Broadcast, UK, English, Main Article, Business

2.3.1 Detection of Article Body

For our research we are interested in collecting the main textual body of an article and not only its summary that is present in the news feed. To find the full text we follow the link in the feed. That page contains much more information than the simple text of the article such as images, links to other pages, advertisements, *etc* (See Fig. 2.1). All this extra information is considered as irrelevant for our purposes and it must be removed before further processing takes place. We developed an HTML Scraper was developed which uses as input an HTML page, removes all the noise and the HTML mark-ups and returns the raw textual body of the article [119, 195]. To build the scraper we implemented a simple strategy of locating the longest chunk of continuous text. This simple approach is the only feasible solution to the problem, since more sophisticated methods based on Machine Learning are too slow to be deployed on the scale that our system requires [44]. The utilised method requires only few milliseconds to scrape a single HTML page compared to the order of seconds that the more sophisticated methods require. The trade-off is the gathering of text chunks that are not guaranteed to include the full textual content of the article.

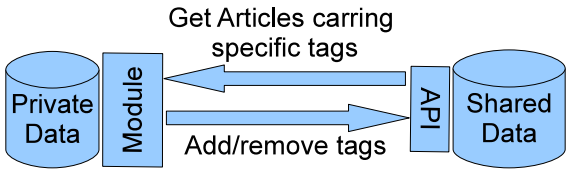


Figure 2.4: Functionality of a generic module.

2.3.2 Tags of News Items

The news items are tagged in multiple ways. First of all they inherit tags from the outlet tag-space and the feed tag-space that carries them. If an article is present in more than one feed then it inherits all the related tags from the corresponding feeds. This can happen for articles that appear, *e.g.*, in the Main feed and in a specialised Topic feed – in this case at least two tags are inherited to the item: one to indicate that the article appeared in the main page of the outlet and one about its topic.

A series of other tags can be applied to a news item in a later processing phase. These tags are applied in an automatic way by our corpus annotation system that is discussed in Sect. 2.4. Example of such tags are the Topics of the item as they are assigned by a Machine Learning method.

2.4 NOAM: News Outlets Analysis & Monitoring System

Our project aims at analysing patterns found across multiple news outlets spread over different countries. Neither the publicly available corpora such as the Reuters, nor the commercial databases such as the LexisNexis provide this required diversity. But, nowadays the content of most news media is available online on their websites.

We built a data management infrastructure capable of gathering online new media content that we refer to as 'News Outlets Analysis & Monitoring' system or NOAM [62]. The system is able to extract information about the

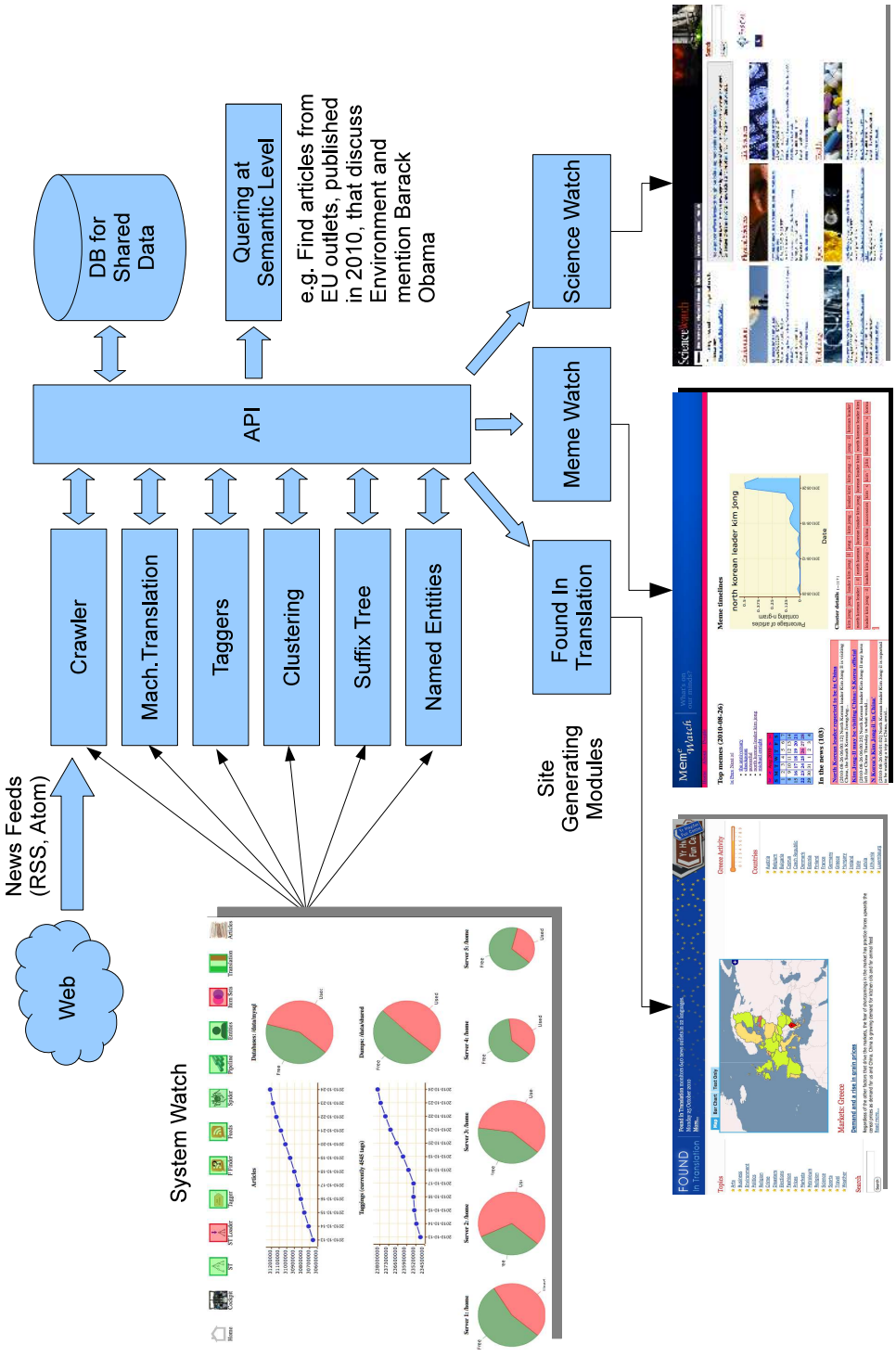


Figure 2.5: NOAM: News Outlets Analysis & Monitoring system.

news at a semantic level and annotate them accordingly. Thus, it allows the querying of the media content at a semantic level and it eases the detection of interesting patterns. NOAM functionality is based on the robust and reliable combination of a several different techniques from the fields of Data Mining, Machine Learning and Natural Language Processing.

2.4.1 Architecture

The architecture of NOAM is highly modular. This allows a distributed and parallel data processing approach, since scalability is a critical requirement of the system. For our design 'modules' are the core elements of the system. Each module implements a different state of the art method of data analysis from Machine Learning, Data Mining and Natural Language Processing. The general functionality of a module is illustrated in Fig. 2.4: it retrieves from the database articles that have some specific set of tags assigned to them; it analyses these articles; and finally it writes back some new annotation to them in the form of tags based on the analysis output. It is worth mentioning that some modules create new content in the form of new articles such as in the case of the Machine Translation module. Modules read and write information to the Database through a common API that guarantees a coherent data access. Finally each module is able to store to the database its own private data which are not shared to the other modules. For example the Toppic Tagger module stores privately the classifier parameters but contributes topic tags to the other modules.

Figure 2.5 presents the overall architecture of NOAM. It is comprised of a series of modules, the main database and the API that enables the access to the DB. Some modules, like the Web Crawler or the Machine Translation module, generate new content. The Output modules generate files in XML format with information that is used by the front-ends or for the creation of reports about the mediasphere content.

Modules work autonomously and independently of each other and in parallel. Even if a module stops working the other modules keep on performing their tasks. By a careful design of a) the tags that modules generate as out-

put and b) the tags that modules require of an article to carry to consider it as input, we can create processing pipelines. An example of such a pipeline is as follows: the Crawler generates new multilingual articles every day; the Machine Translation module translates the non-English articles into English; the Topic Taggers generate new annotations; and finally an Output module generates the desired report of the topic distribution of the leading articles of the day.

Fig. 2.6 illustrates the Entity-Relationship model of the main entities of the DB behind NOAM. The main entities are the 'ARTICLE' that stores articles; 'OUTLET' that stores the different outlets;'FEED' that stores the different Feeds; and 'STORY' that groups articles into stories (Stories are explained next in Sect. 2.5). The first three entities are coming with an extra table, one for each entity, that stores the corresponding tags for that entity. There is a many-to-many relation between tags and the corresponding Entities.

2.4.2 Implemented Modules

In this section we summarize the main modules that currently make up the NOAM system and we emphasize which ones were not part of this thesis-work:

- *Crawler.* It craws a predefined list of news feeds of interest. The feeds can be in either Atom or RSS format. After collecting information such as Title, Description, and publication date for each article, the Crawler tags them with a set of tags that will enable other modules to work on them, for example the language of article (See Sect. 2.2[3]).

- *HTML scraper.* It collects the main body of the article based on the link provided by the news-feed (See Sect. 2.3.1).

- *Feed Finder.* It is responsible for discovering new feeds based on some manually set starting points. The approval and the addition of the

[3]This module was implemented by Omar Ali.

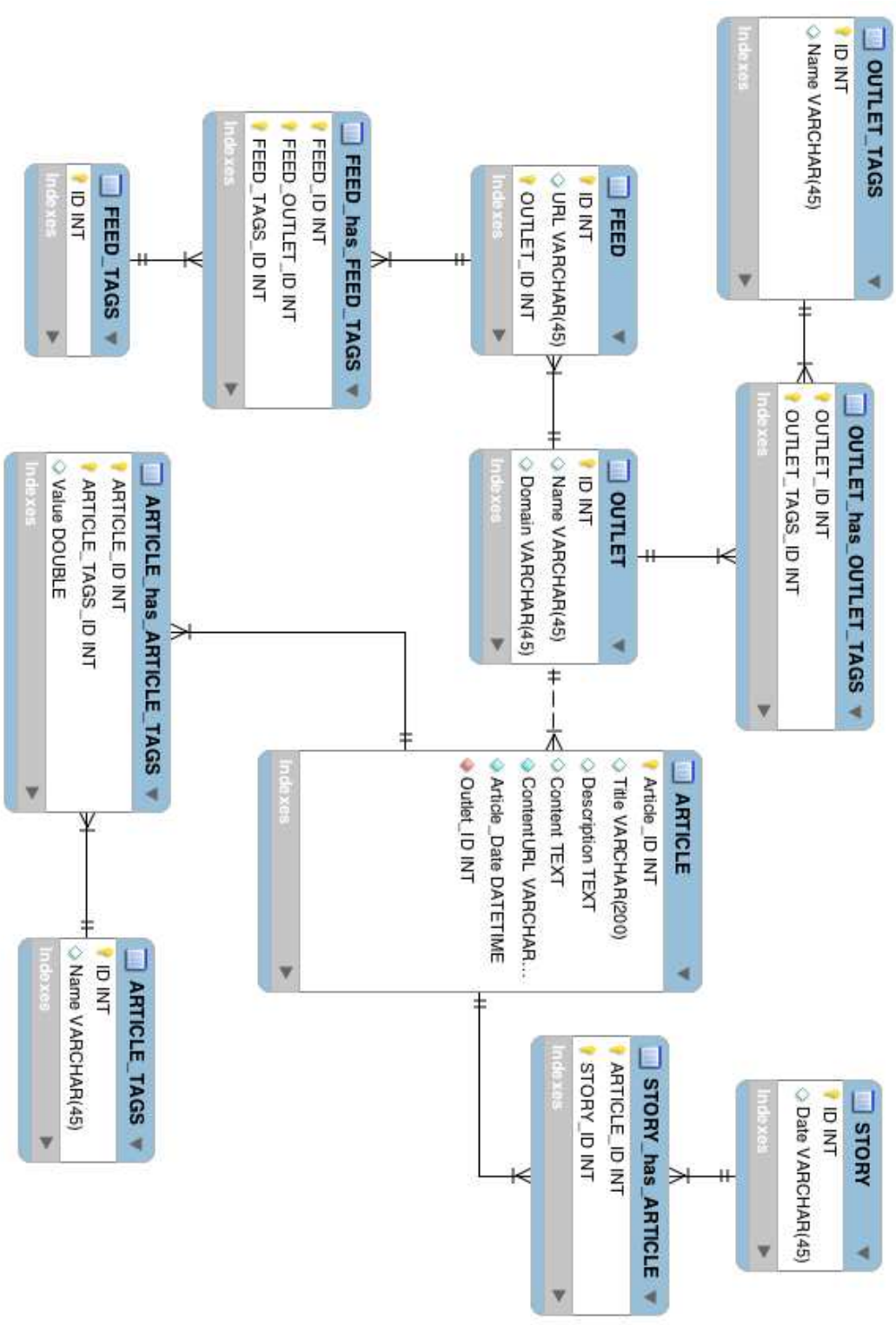


Figure 2.6: Entity-Relationship model of NOAM.

new feeds into the system is performed in a semi-automatic way (See Sect. 2.2.1).

- *Machine Translation.* It translates all non-English articles into English using a phrased based statistical machine translation approach[4] (See Sect. 2.4.3) [182].

- *Topic Taggers.* It discovers the general topics of each article using topic classification based on Support Vector Machines (See Chap. 3).

- *Clustering.* It clusters articles into stories, that is sets of articles that discuss the same event (See Sect. 2.5).

- *Suffix Tree.* It implements a suffix tree that enables the detection of frequent phrases and text reuse[5] [165].

- *Entities Detection.* It detects Named Entities, such as people, organisations and locations, present in the articles[6] [5]. Their detection is based on GATE [40].

- *Output Modules.* A set of modules that generate outputs in XML format which are used by the NOAM front-ends such as 'Found In Translation' and 'Science Watch' (See Sect. 3.3). They are activated typically once per day producing reports based on the data of the previous day.

NOAM is equipped with a back-end system monitoring tool, namely System Watch. It is a web based graphical user interface that enables an overall view of the system status, such as information about the status of the physical servers where the modules and DB live. It also enables the control of most modules of the NOAM system. For example, it enables the tuning of taggers parameters; the importing of new outlets and feeds into the Crawler (illustrated in Fig. 8.5); the annotation of Feeds with tags (illustrated in Fig. 8.4); the query of the database for specific articles (illustrated in Fig. 8.3).

[4]This module was implemented by Marco Turchi.
[5]This module was implemented by Tristan Snowsill
[6]This module was implemented by Omar Ali.

Table 2.4: Machine Translated articles per language (February 11th, 2011).

Language	Articles	Language	Articles
Bulgarian	153535	Latvian	87682
Czech	276455	Lithuanian	248503
Danish	676399	Maltese	4212
Dutch	951851	Polish	380044
Estonian	254800	Portuguese	299538
Finnish	504159	Romanian	316135
French	604073	Slovak	166809
German	898498	Slovenian	214383
Greek	401606	Spanish	1314433
Hungarian	195026	Swedish	766481
Italian	715446		

2.4.3 Machine Translation

We applied a Statistical Machine Translation (SMT) approach for translating the non-English articles to English[7]. SMT is based on a noisy channel model [20], where a Markovian Language Model coupled with a phrase-to-phrase translation table are at the heart. In recent years, the noisy channel model has been extended in different directions. The most fruitful has been the phrase based statistical machine translation (PB-SMT) introduced by Koehn et al. [106] that is based on the use of phrases rather than words. We use Moses, a complete phrase based translation toolkit for academic purposes. It provides all the state of the art components needed to create a PB-SMT system from one language to another. For each language pair, an instance of Moses is trained using either Europarl [105] data or the JRC-Acquis Multilingual Parallel Corpus [170].

We translated all non-English articles of 21 official EU languages into English. The number of translated articles per language is illustrated in Table 2.4.

We make the working assumption that SMT does not alter significantly the geometry of the corpus in the vector-space representation of articles

[7]The development of the Machine Translation module was made by Marco Turchi.

(for vector spaces see Sect. 2.4.4). This is corroborated by results on cross-language information-retrieval, where vector space representations of translated documents are successfully used [158].

2.4.4 Text Pre-processing

Before the textual content of articles, such as title, description and content, is processed by modules such as clustering and classification some preprocessing steps are required. Usually these include stop-word removal, stemming, handling of punctuation, cases of letters *etc*.

Stop-word removal

A lot of words that appear in text like articles and prepositions carry little or no information [119]. These are called stop words and in terms of pattern analysis can be considered as noise. The detection of the words is based on publicly available lists. Examples of such words in English language include: *the, and, in, on, at, what, will, with, etc*. These words are removed.

Stemming

In many languages, including English, words appear in different forms depending on grammatical and syntactical rules. For example, nouns have plural forms and verbs used in past tense have a different form to the present tense. For purposes of tagging or clustering of texts the different forms of the same words is an issue. A common method used that partially solves the problem is stemming [119].

Stemming is the process of reducing words to their roots (stems). For English language this process is mainly the removal of the suffixes of the words. For example the words 'computers', 'compute' and 'computing' are all reduced to the same stem 'comput'. Several algorithms have been proposed for stemming. We adopted one of the most popular used in literature that is developed by Porter and which is based on a set of specific rules [150].

Text representation

The pre-processed document is represented in the form known as *Vector Space Model* [156]. Each word is considered a term. Term sequence and position in the sentence or in the document are ignored. So, given a set of documents $D = \{x_1, x_2, ..x_m\}$ we define $V = \{t_1, t_2, ...t_n\}$ be the set of distinctive terms t_i found in D. Since we stemmed the documents, the terms correspond to the stemmed formed of the words. V is known as the vocabulary. In order to represent a document $x_j \in D$ we assign a weight $w_{ij} \geq 0$ to each term $t_i \in V$. So each document can be represented with a term vector $x_j = (w_{1j}, w_{2j}, ...w_{nj})$ where w_{ij} correspond to the weight of term t_i found in document x_j.

Several different models have been proposed for calculating w_{ij}. In boolean model w_{ij} is simply 1 if the term t_i is present in document x_j and 0 otherwise. A more advanced representation is the TF-IDF scheme which we adopted for our research [156, 119]. It is based on calculating the term frequencies (TF) and the Inverse Document frequency (IDF). TF of term t_i in document x_j is calculated by

$$tf(t_i, x_j) = \frac{f_{ij}}{\sum_k f_{kj}} \tag{2.1}$$

where f_{ij} is the number of occurrences of the considered term in document x_j and the denominator is the sum of occurrences of all terms in document x_j.

The IDF of term t_i is calculated by

$$idf(t_i) = \log \frac{m}{|x : t_i \in x|} \tag{2.2}$$

where m is the number of documents in corpus and the denominator is the number of documents where the term t_i appears.

The TF-IDF weight of term t_i in document x_j is defined as the quantity

$$tfidf(t_i, x_j) = tf(t_i, x_j) \cdot idf(t_i)$$

The physical meaning of TF-IDF is the assignment of a high weight to terms

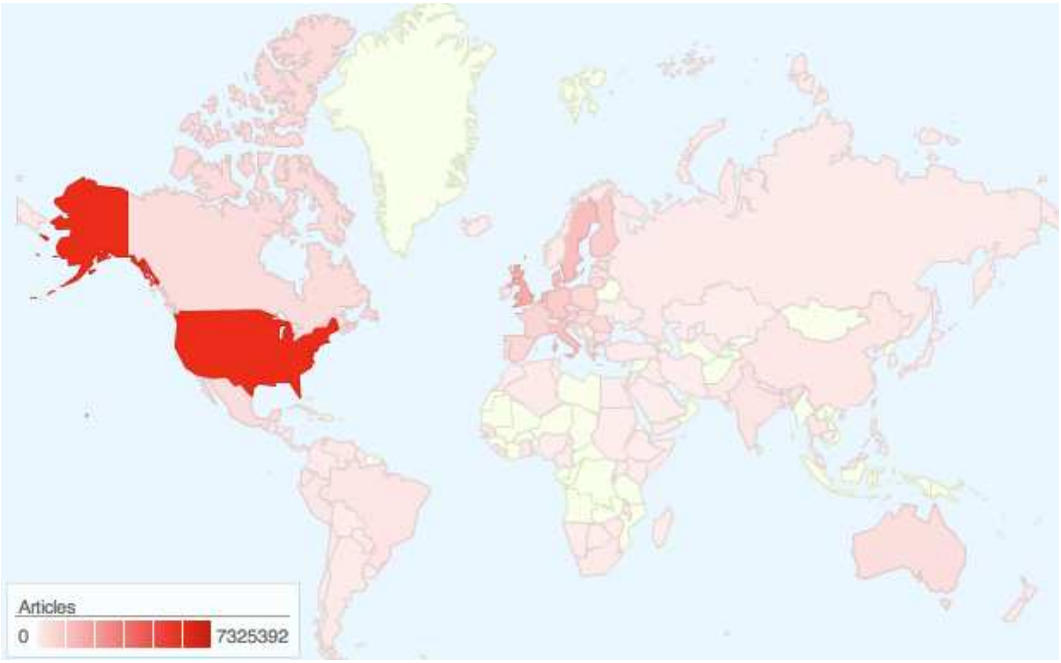


Figure 2.7: Map visualising number of articles by country analysed by NOAM as on February 11th, 2011.

that appear many times in a document but are not present in the majority of documents.

2.4.5 Our Corpus

Currently we monitor more than 1772 news outlets and 3888 news feeds. The full list of outlets and number of feeds per outlet tracked by NOAM is illustrated in Supplementary Table 8.2. These outlets are from 193 distinct countries – or wider geographic regions, such as 'EU' or 'Latin American'. Table 8.1 presents the number of outlets and feeds we track per geographic region, while Fig. 2.7 offers a spatial visualisation of the number of collected articles so far per country. We track news in 21 different non-English languages all of which are machine translated into English. The number of outlets and feeds we track per language is illustrated in Table 2.6. We divide outlets in 8 media types as illustrated in Table 2.5. So far, we have analysed more than 30 million news items, and currently we collect and analyse more

Table 2.5: Outlets and Feeds per Type in NOAM (March 8th, 2011)

Type	#Outlets	#Feeds
Blog	169	233
Broadcast	88	388
Magazine	88	324
News Community	15	54
Newspaper	904	1673
Newswire	33	104
Online Only Media	134	188
Press Releases	263	382

Table 2.6: Number of Outlets and Feeds in NOAM by Language

Language	#Outlets	#Feeds	Language	#Outlets	#Feeds
Bulgarian	14	14	Italian	34	76
Czech	10	12	Latvian	7	7
Danish	22	22	Lithuanian	11	11
Dutch	42	43	Maltese	4	5
English	1119	2563	Polish	24	24
Estonian	12	12	Portuguese	30	37
Finnish	33	33	Romanian	16	16
French	56	82	Slovak	8	8
German	52	83	Slovenian	6	7
Greek	34	48	Spanish	120	182
Hungarian	15	18	Swedish	44	48

than 40K new news items per day.

The volume of articles that are published is not constant day by day. We found that it follows a significant periodic behaviour with a seven days cycle. To illustrate this behaviour we run two tests on two independent subsets of outlets, the English and the non-English ones. Figure 2.8 presents a time-series of the articles we collect per day from English language media and Fig. 2.11 present the same period time-series for non-English media that were machine translated into English. Figures 2.10 and Fig. 2.13 illustrate the corresponding spectral power distributions. We can observe the seven days period in the volume of news articles published. During weekends there is a significant reduction of the news items that media publish compared to

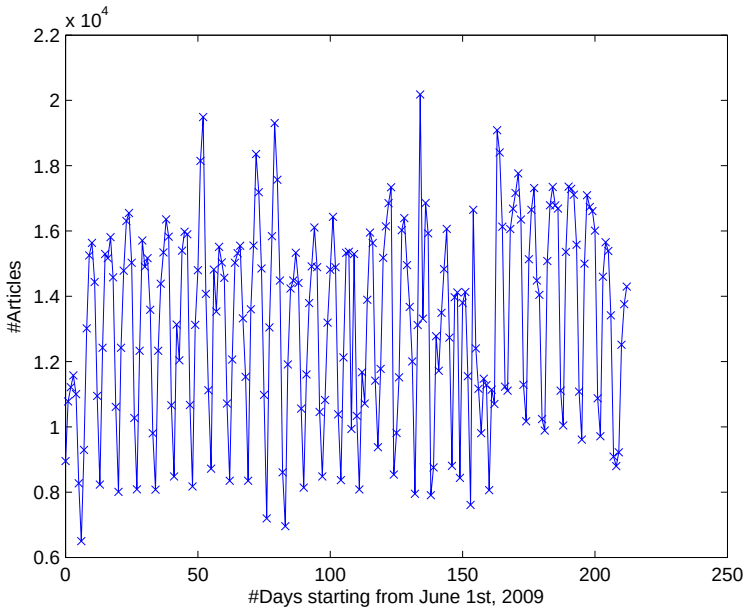


Figure 2.8: Number of English articles per day.

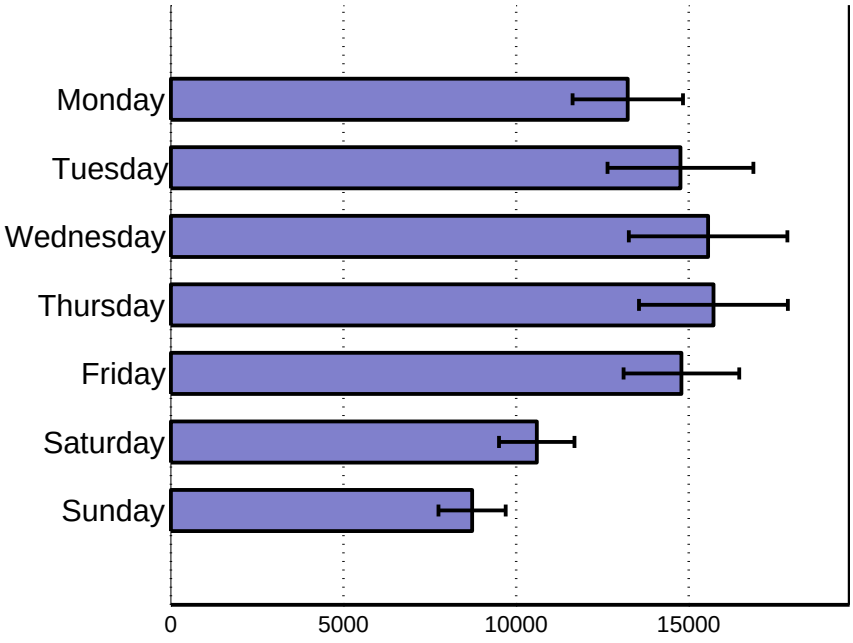


Figure 2.9: Average number of English articles per day of week. Errorbars are Standard Deviation.

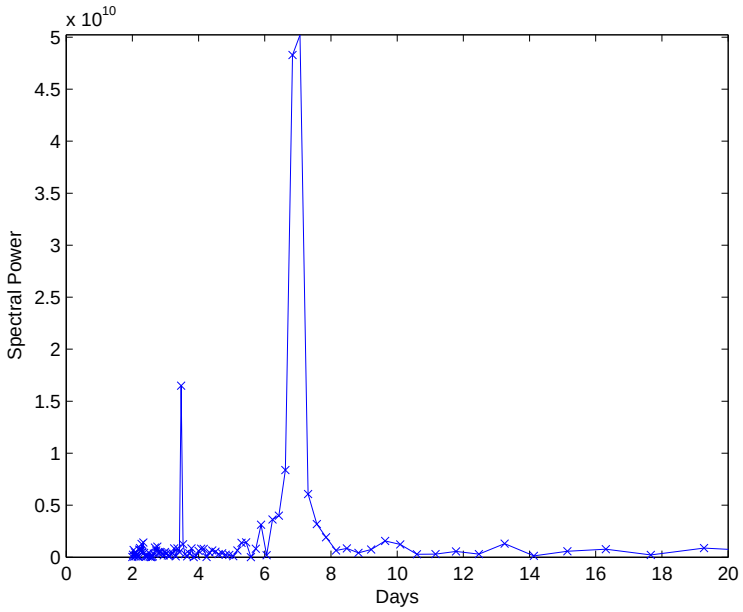


Figure 2.10: The seven days cycle on the volume of published articles in English. Plot illustrates the spectral power distribution of the corresponding time-series illustrated in Fig. 2.8.

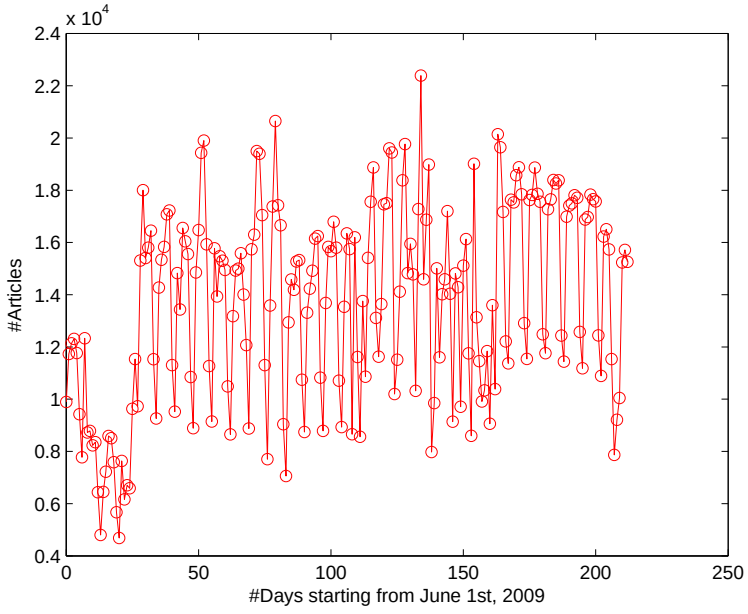


Figure 2.11: Number of Machine Translated articles per day.

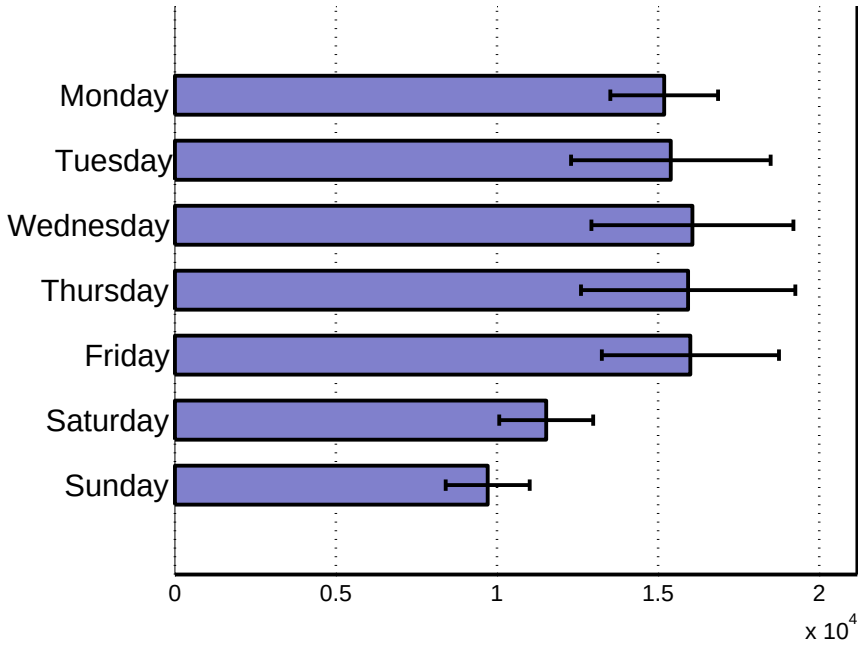


Figure 2.12: Average number of Machine Translated articles per day of week. Errorbars are Standard Deviation.

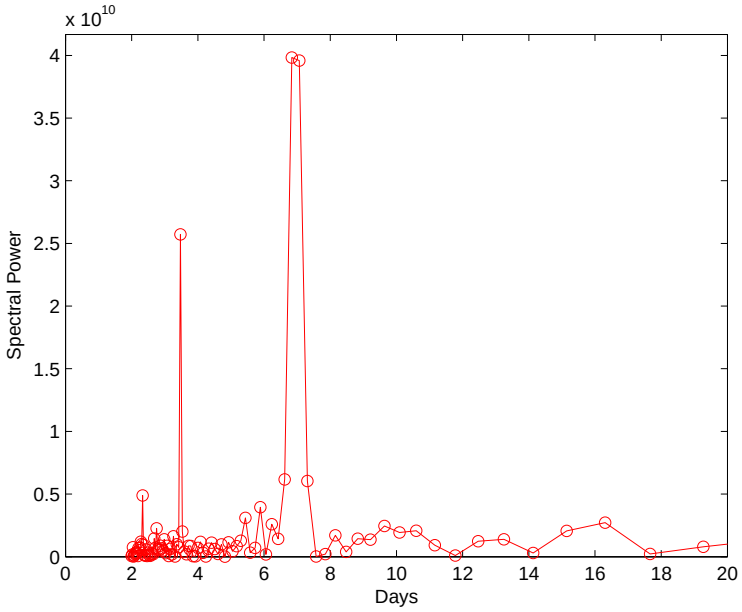


Figure 2.13: The seven days cycle on the volume of published articles in languages other than English. Plot illustrates the spectral power distribution of the corresponding time-series illustrated in Fig. 2.11.

the volume at the middle of the week as illustrated in Fig. 2.9 and Fig. 2.12. This "discontinuous" behaviour explains the 3.5 days cycles that appear in Fig. 2.10 and in Fig. 2.13. Similar to our findings is the seven days periodic behaviour that has been found in blogs content [113].

2.5 From Articles to Stories

We define a news story as a set of news items that discuss the same event. To detect the stories that are present in a set of news items we need to achieve two things. Firstly to define a similarity measure between news items, and secondly to apply a clustering method based on that similarity measure. Clustering is one of the most widely used methods for data analysis. It is used in many different fields where an understanding of the patterns present in data is required. A formal definition of clustering is quite difficult and application dependent since there is no ground truth for what is considered a good clustering of data [123]. In this research a cluster is defined as a set of news articles that cover the same event. We are interested in forming them, representing them, and in extracting information out of them. In our research we adopted the well accepted cosine similarity measure [119], and we applied the Best Reciprocal Hit clustering method.

The cosine similarity K of two documents x_i and x_j is given by:

$$K(x_i, x_j) = \frac{\langle x_i, x_j \rangle}{\| x_i \| \cdot \| x_j \|}$$

where x_i and x_j are news items in the bag-of-words vector space, and $\| \cdot \|$ refer to the Euclidean norm of the vector. It ranges from zero to one and represent the angle between two vectors. It is similar to the dot product of the two vectors, but it is normalised to their length. This is very useful in text mining since texts are usually of different lengths (in words).

We cluster news articles using the Best Reciprocal Hit (BRH) method, borrowed from the field of bioinformatics [91, 101]. The advantages of this clustering method for the current problem are that a) we don't need to specify the number of clusters we want to discover since it is unknown and non-

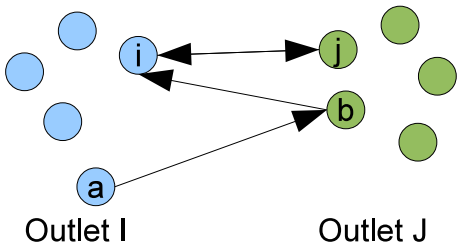


Figure 2.14: Example of usage of the BRH algorithm. Circles represent articles and arrows connect the closest articles.

stationary per day and b) we take advantage of the natural separation of the articles into sets of articles published in different news outlets. The BRH method has been used in the past in news analysis research for the discovery of similar articles between different news outlets [68].

In proteomics a protein i in genome I is a BRH of protein j in genome J if query of genome J with protein i yields as the top hit protein j, and reciprocal query of genome I with protein j yields as the top hit protein i. Analogous in the case of media analysis is that an outlet corresponds to a genome and an article to a protein. Thus, an article x_i in outlet I is a BRH of article x_j in outlet J if query of outlet J with article x_i yields as the top hit article x_j, and the reciprocal query of outlet I with article x_j yields as the top hit article x_i. For more clarity, this is illustrated in Fig. 2.14: Article x_i is BRH of article x_j, but x_a is not BRH of x_b since the former is closer to x_c. This process results in pairs of articles from different outlets that publish the same stories. Connections of these pairs in larger components is based on the logic that friends of friends are also friends, *i.e.*, the pairs of articles $\{x_i, x_j\}$ and $\{x_j, x_k\}$ form the cluster $\{x_i, x_j, x_k\}$.

The evaluation of the clustering method is a difficult task. We have monitored the output of the BRH algorithm over several days and the inspection of the results showed sensible clusters of articles that referred to the same event. Furthermore, it is worth noting that the analysis of relations of outlets and countries, that we will present in Chapter 6, it is based on clusters created using the BRH method, and it leads to significant results that are

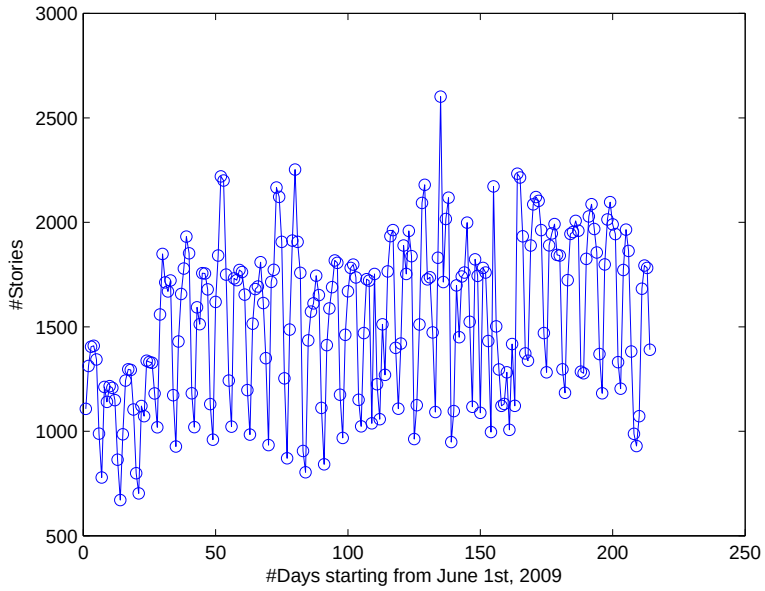


Figure 2.15: Timeline of stories per day.

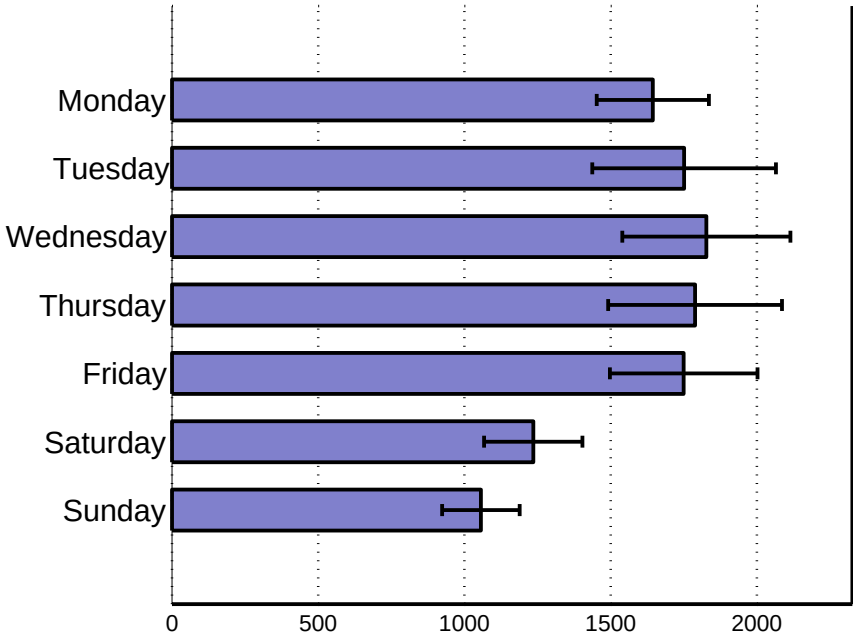


Figure 2.16: Average number of stories per day of week.

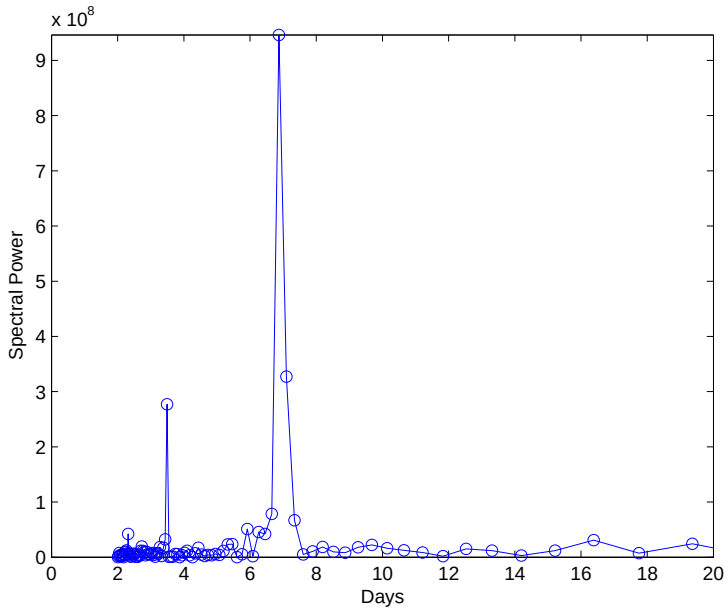


Figure 2.17: The seven days cycle on the volume of published stories. Plot illustrates the spectral power distribution of the number of stories published per day.

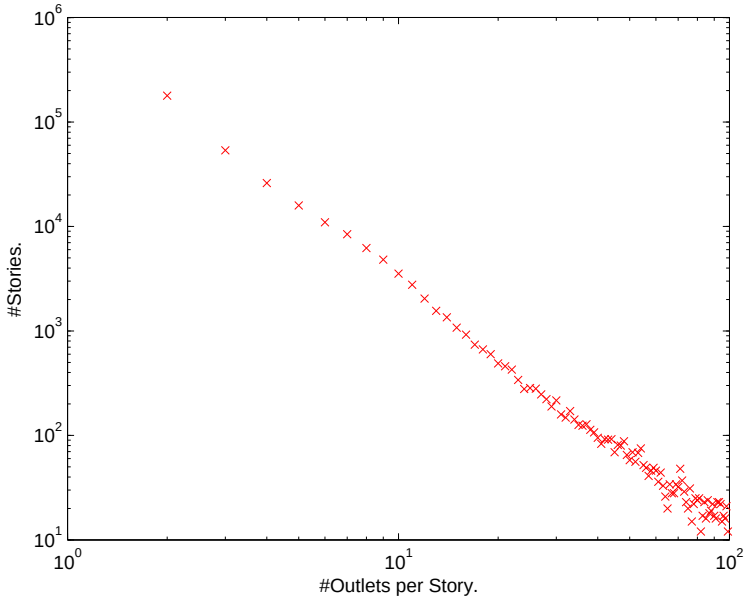


Figure 2.18: Number of outlets per story.

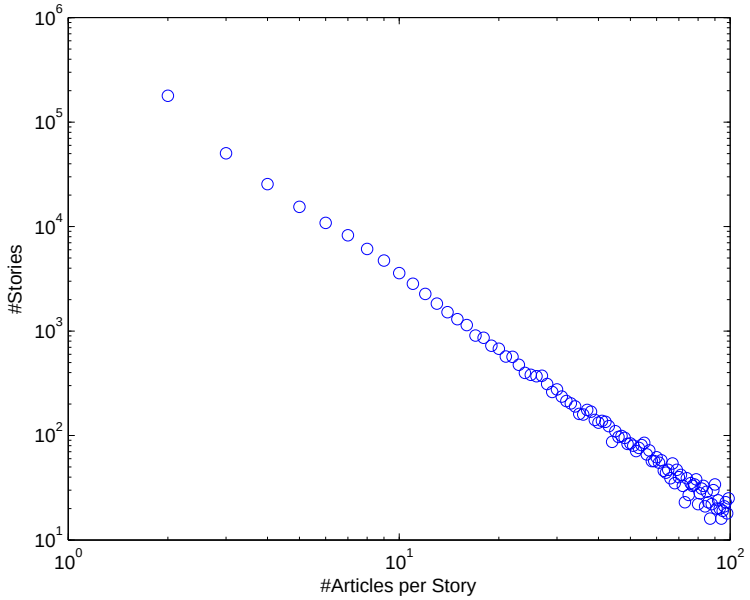


Figure 2.19: Number of articles per story.

correlated to ground truth data.

Figure 2.15 illustrates a timeline of the stories per day that were detected with our system. In Fig. 2.17 we plot the spectral power distribution of the aforementioned signal. We can observe a periodic behaviour of seven days cycle, similar to the case of the articles per day. As in the case of published articles per day there is a significant lower number of published stories during the weekends, illustrated in Fig. 2.16. This behaviour explains the 3.5 days cycle artefact of Fig. 2.17.

Figure 2.18 presents the number of stories as a function of the number of outlets that cover the same story. It can be seen that few stories are covered by hundreds of outlets, while thousands of stories are covered by very few media. The plot is in log-log space and it indicates a power law distribution. By applying the power-law distribution fitting methodology for empirical data of Clauset *et al.* [36] we discovered that data are distributed based on the power law of the form $x^{-\alpha}$ with $\alpha = 0.67$. A similar power law is present for the case of articles per story with $\alpha = 0.60$, as illustrated in Fig. 2.19. The physical meaning of this power law is that there are few stories

comprised of many articles and many stories comprised of few articles.

2.6 Summary

In this chapter we introduced the notion of news outlets, feeds and articles that we use throughout the thesis, and discussed laws that dominate the media system. More precisely:

- we presented NOAM the system that was used for the gathering and annotation of news articles. Currently we analyse 40K news items per day and so far we have analysed more than 30 million news items and we monitor
  - more than a 1772 news outlets
  - more than 3888 news feeds
  - in 22 different languages
  - from 193 distinct geographic regions
  - from 8 media types
- we showed a significant seven days periodic behaviour to the number of articles published
- we clustered articles into stories using the Best Reciprocal Hit method.
- we showed a similar seven days periodic behaviour to the number of stories published
- we showed that the number of outlets that cover a story follows a power-law.

In the next chapter, we will use Support Vector Machines to automatically detect the topics of the articles, a process similar to the 'coding' that is conducted by media scholars for the analysis of news. In the rest of the research we will use several well defined parts of this annotated corpus to analyse the media system and detect the emerging patterns.

Chapter 3

Machine Learning for News Analysis

In this chapter we present Machine Learning (ML) methods that can be applied for the annotation, or *tagging*, of news media content. This is towards the automation of the manual news analysis method performed by social scientists known as 'coding'. We deploy kernel methods and in particular Support Vector Machines (SVMs) for the automatic annotation of news items based on their topic. We build a series of taggers based on the Reuters, The New York Times as well as our own corpus. We validated and tuned them by measuring their precision and recall on unseen test-sets of articles. We adjust the trade-off between precision and recall to fit our needs and annotate our corpus accordingly. Next, we used them to develop two demos that demonstrate the ability of constant media monitoring and news annotation. Finally, we show that using a SVM-ranking approach we can predict which news items are more appealing to a given audience and furthermore, we generate lists of keywords that are expected to trigger the attention of readers.

3.1 Kernel Methods

We base our analysis on Machine Learning methods derived from the class of techniques known as *Kernel Methods* [164, 159]. These methods include Support Vector Machines and they are based on a pairwise similarity between data points, or more specifically for our research, similarity between news items.

In general, let Φ be a non-linear mapping from the input space X to the feature space $F \in H$ where H is a Hilbert space. The function $k(x_i, x_j)$, known as *kernel function*, is defined as

$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_H \tag{3.1}$$

where $\langle , \rangle_H$ is the inner product of data samples x_i and x_j in Hilbert space H. A *kernel matrix*, also known as *Gram matrix*, is a square matrix $K \in R^{n \times n}$ such that $K_{ij} = k(x_i, x_j)$ for some $x_1, ..., x_n \in X$ and some kernel function k. A Kernel matrix is a symmetric, positive semi-definite matrix, and it completely determines the relative positions of the data points in the embedding space.

In the case of text classification due to the high dimensionality of the features space, which is equal to the number of different terms, a simple linear kernel or its normalized form of cosine kernel, can perform adequately well [176].

3.1.1 Two-Class SVMs

SVMs are kernel-based machine-learning algorithms derived by Vapnik et al. [15] in the framework of structural risk minimisation [186, 26, 39]. Consider an input space X and some training vectors $x_i, i = 1, 2, \ldots l$, belonging to two different classes denoted as $y_i \in \{-1, +1\}$.

SVMs maximize the margin $1/||w||_2$ that separates the two classes by finding the hyperplane (w, β):

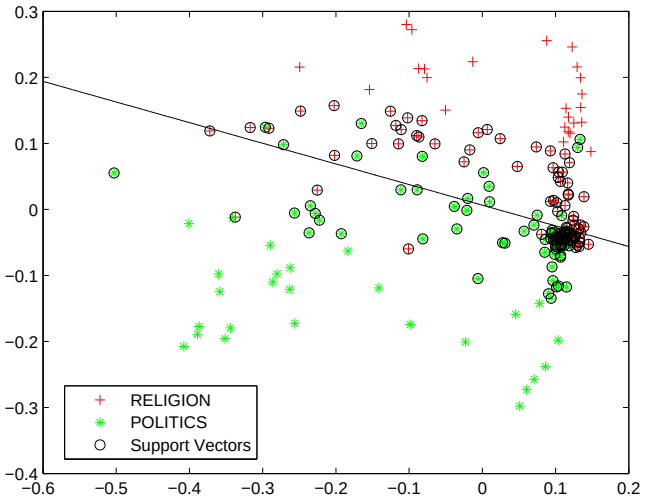


Figure 3.1: Example of SVMs for the separation of news articles of two topics, namely Religion and Politics. The articles are projected on a 2D plane for visualisation purposes using MDS. The line is the result of the SVM that separates samples in that 2D space.

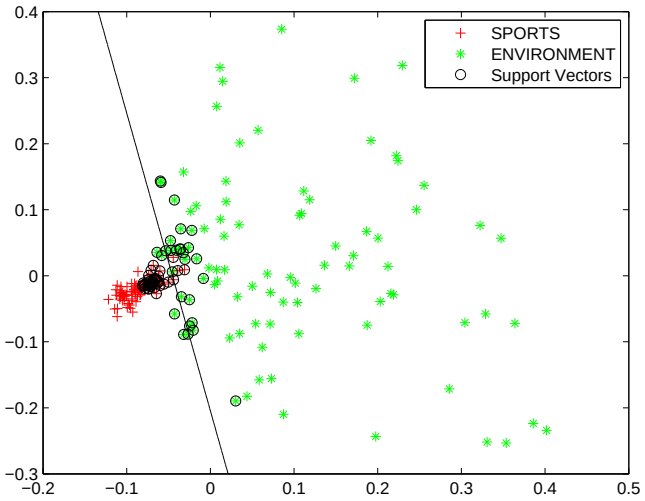


Figure 3.2: The SVM classifier that separates samples of Sports vs. Enviroment topics.

$$\begin{aligned} \text{minimise}_{\xi,w,\beta} \quad & \langle w,w\rangle + C\sum_{i=1}^{l}\xi_i, \\ \text{subject to} \quad & y_i(\langle w,x_i\rangle) \geq 1-\xi_i, i=1,\ldots,l \\ & \xi_k \geq 0, i=1,\ldots,l \end{aligned} \tag{3.2}$$

where ξ_i are positive slack variables that relax the violated constraints in he optimization for non-separable classes, the constraints denote that each vector x_i will be on the same side as the other vectors of the same class and C is a positive parameter which tunes the trade-off between errors and margin width.

This problem is solved by calculating the Lagrangian:

$$L(w,\beta,\xi,\alpha) = \frac{1}{2}\langle w,w\rangle + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i[y_i(\langle w,x_i\rangle+\beta)-1+\xi_i]$$

The Lagrangian dual is calculated by:

$$L(\beta,\xi,\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle - \frac{1}{2C}\langle\alpha,\alpha\rangle \tag{3.3}$$

We observe that the relations of x_i and x_j are only threw their inner products. Thus, we can use the kernel trick and replace the inner product with a desired kernel function $k(x_i,x_j)$. The quadratic problem is solved by numerical approaches such as the Sequential Minimal Optimisation (SMO) [149].

Two open-source implementations of SVMs, namely LibSVM [33] and SVM^{Light} [97] were used for the purposes of the project. Both of them use similar data structures and implementation algorithms. Optimized versions of these programs are also available for the case of use of only linear kernels, namely LibLinear [56] and SVM-perf [99] respectively.

Example of topic classification using SVMs

As an example, we illustrate the two-class SVMs in action for the separation of articles belonging to two different topics. The articles are represented by high dimensional vectors, one dimension for each word in the vocabulary. To visualise the articles in a 2D plane we utilise a dimensionality reduction technique known as Multidimensional Scaling (MDS) [107, 37].

MDS is a data visualisation technique aiming to embedding data points $x_1, ...x_n \in \mathbf{R}^N$ into some low dimensional space. A $\mathbf{R}^2$ or $\mathbf{R}^3$ space is suitable for visualisation. The MDS algorithm tries to find the best positions of the points in the low-dimensional space so as to keep their relative positions as close as possible to their real positions in the high dimensional space. The input of the MDS algorithm is a matrix of all distances $\delta_{ij} = \|x_i - x_j\|$ between all pairs of points x_i and x_j, and $\|\cdot\|$ denotes a norm in that space. There are many different flavours of the MDS algorithm based on the selection of the distance (or the similarity) that is used. For a visualisation in a 2D plane, MDS algorithm tries to find the best points $\hat{x}_1, ..., \hat{x}_n \in \mathbf{R^2}$ such that $\|\hat{x}_i - \hat{x}_j\| \approx \delta_{ij}$. This leads to a minimization problem of a cost function of the form

$$\min_{\hat{x}_1,..\hat{x}_n} \sum_{i,j} \left(\|\hat{x}_i - \hat{x}_j\| - \delta_{ij}\right)^2 \tag{3.4}$$

For the visualisation of our examples we used as input of MDS the cosine kernel matrix of 100 randomly selected articles per class from labelled data and requested a 2D output. Then a linear SVM separates them on the 2D plane. In Fig. 3.1 we separate articles of Politics from articles of Religion, while on Fig. 3.2 we separate articles of Sports from articles of Environment. In both examples we also plot the hyperplane w, or in this case the straight line, that separates the topics. We also indicate the points that have a non-zero α which define that line and are known as support vectors. The two visualisations also provide some insight of the geometry of the articles space on which we will work on for the rest of the chapter.

3.1.2 One-Class SVMs

A difficulty in news classification is the definition of the negative class of samples, *e.g.*, What is non-Politics? A solution to tackle this problem was proposed by Scholkopf et al. [160] and it is based on SVMs trained using only positive samples of data. This approach, namely one-class SVM, removes the need of the definition of the negative class. The trade-off is an expected decrease in accuracy [125, 96, 46].

We checked this approach for the purposes of our project but unfortunately the results were quite poor for both linear and cosine kernels close to random guessing, as reported in, *e.g.*, [125]. We based our experiments on LibSVM which implements the one-class SVMs. A recent study showed that by utilising both positive and unlabelled examples, and a series of feature pre-processing techniques such as normalising for document length, it is possible to reach only up to 95% of the performance of two-class SVMs [191]. Thus, for our project we adopted the two-class SVMs. The typical solution of the negative class definition problem is the one-vs-all strategy, that is to consider a random sample of non-positive documents and treating them as the samples of the negative class. In other words, combine the non-positive classes into a single super-class and use it as the negative class.

3.1.3 Ranking with SVMs

SVMs have been extended to deal with the task of ranking items [97]. This learning process is based on pairwise preference relations between pairs (x_i, x_j) of data. If we denote that item x_i is preferred to x_j by $x_i \succ x_j$ and we define a linear function $u : \mathbb{R}^n \to \mathbb{R}$ of the form $\langle w, x \rangle$, we can express their relationship as:

$$x_i \succ x_j \iff u(x_i) > u(x_j) \iff \langle w, (x_i - x_j) \rangle > 0 \tag{3.5}$$

We can express the relationship learning of l items x_i and x_j as a binary classification task by denoting their differences by $s_k \in S$, with S the set of

all pairwise differences, and solving the quadratic optimisation problem:

$$\begin{aligned} \text{minimise}_{\xi,w} \quad & \langle w, w \rangle + C \sum_{k=1}^{l} \xi_k \\ \text{subject to} \quad & y_k(\langle w, s_k \rangle) \geq 1 - \xi_k, \\ & \xi_k \geq 0 \quad \forall k = 1, \ldots, l \end{aligned} \tag{3.6}$$

where ξ_k are the slack variables that allow us to deal with non-separable data. The w vector is the desired solution that can provide the ranking of two items by assigning class label $y_k = +1$ if $\langle w, s_k \rangle \geq 0$, and $y_k = -1$ otherwise.

3.1.4 Performance Measures

The empirical evaluation of the quality of a classifier can be realised using a test set with known labels. We try to predict the real label $y = \{+1, -1\}$ of each sample with the classifier, which outputs its prediction $\hat{y} = \{+1, -1\}$. For each such prediction there are four possible outcomes that are named as:

- True-Positive (TP), if $y = +1$ and $\hat{y} = +1$,
- False-Positive (FP), if $y = -1$ and $\hat{y} = +1$,
- False-Negative(FN), if $y = +1$ and $\hat{y} = -1$,
- True-Negative (TN), if $y = -1$ and $\hat{y} = -1$,

Given a test set of N labelled samples we can define a series of performance measures for a classifier such as Accuracy, Precision, Recall and F_β-Score:

- *Accuracy*

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

  Accuracy is a measure that can easily be misleading. It is informative only in the case of equally distributed samples among the two classes, *e.g.*, if most test samples belong to the negative class, and the classifier always predicts the negative label independently of the sample, then

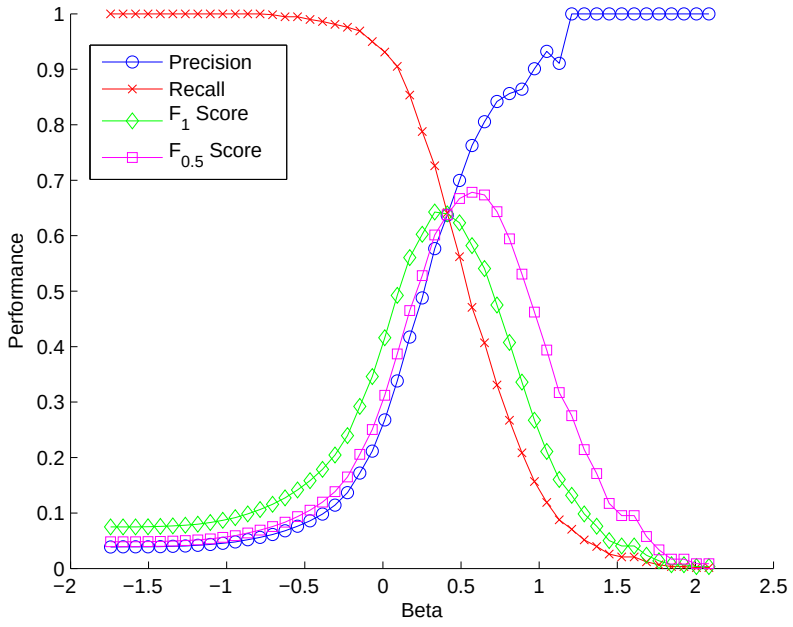


Figure 3.3: Example of Precision, Recall and F_β curves for different SVM decision thresholds for the 'Crime' tagger.

its accuracy will appear high and equal to the proportion of negative to positive samples in the test set.

- *Precision & Recall*

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Precision and Recall are two complementary and easily interpretable measures of the performance of a classifier. Precision reveals how accurate are the predictions of the positive samples of the classifier. Recall measures how many of the positives samples were discovered. The ideal classifier should have both high precision and recall, that is it must be able to find all positive samples without errors.

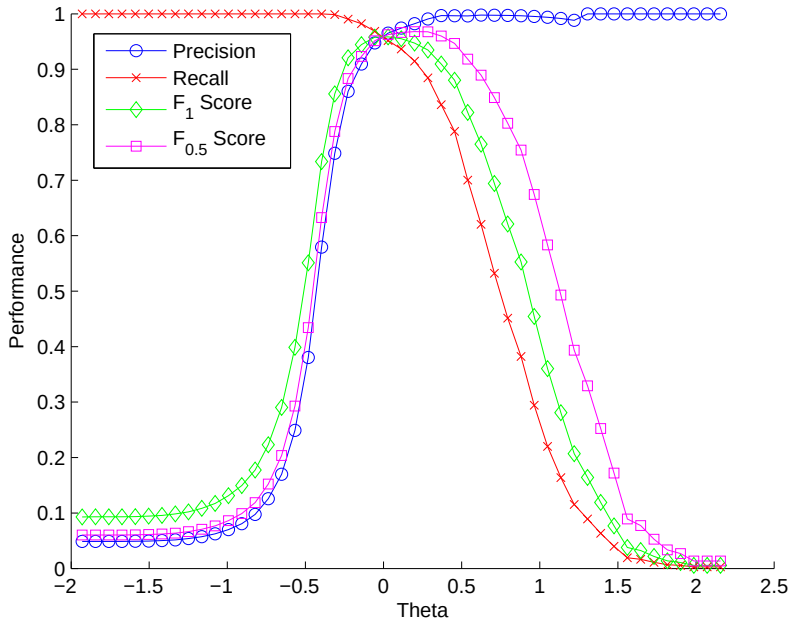


Figure 3.4: Example of Precision, Recall and F_β curves for different SVM decision thresholds for the 'Sport' tagger.

- *F_β-Score*

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

  This scores combines the Precision and Recall in a single measurement. The β parameter defines a weight among Precision and Recall. Which one is more important, Precision or Recall, depends on the problem under study. For $\beta = 1$ we get the F_1 score that treats Precision with the same weight as Recall, and for $\beta = 0.5$ we get the $F_{0.5}$ that treats Precision as twice as important as Recall.

The decision threshold of the SVM can be adjusted in order to achieve the required trade off between precision and recall [162]. This allows the deployment a flexible set of taggers that can work, *e.g.*, on either the high-Precision area of the spectrum or, by just adjusting the threshold, on, *e.g.*, the optimum $F_{0.5}$ score. Figure 3.3 illustrates an example of the different Precision, Recall, F_1 and $F_{0.5}$ curves that are created for different decision

thresholds of the classifier for the detection of news items about 'Crime' articles. Figure 3.4 illustrates a similar example for the 'Sports' classifier.

Other more sophisticated methods for evaluating taggers performance include the Area Under the Received Operating Characteristic (ROC) Curve (AUC). The AUC allows the evaluation of taggers for all different decisions thresholds [57, 61, 58]. For the evaluation of taggers we adopted the simpler aforementioned methods of Precision, Recall and F_β-Score and we will use the AUC approach in Chap. 5 were we compare network inference algorithms and the predictability of the structure of the inferred networks.

3.2 Experiments

3.2.1 Tagging of News Content

We deploy binary SVMs for the categorisation of news items based on their topic. This process is also referred to as tagging since each topic is represented by a tag on the news item. We will use a set of SVM classifiers, one per tag we want to detect. We allow the tagging of an article with multiple tags, in a manner similar to the 'coding' process of social scientists.

The advantages of using binary SVMs for text classification can be summarised as follows: they provide good and robust generalisation performance, outperforming other conventional learning methods for the task of text classification [192, 39, 46, 161, 97]. They are computationally efficient, since by taking advantage of duality they do not depend on the dimensionality of feature space but on the size of training set [39]. In text classification the size of feature space is orders of magnitude greater than the size of available samples, but they resist the curse of dimensionality [181]; SVMs are quite robust at handling high dimensionality feature spaces so that no feature selection [81] is necessary to increase performance [97]. It is easy to correlate the learning theory that describe SVM mechanics with respect to text classification.

We trained a series of taggers based on data from the Reuters and the New York Times corpora, as well as from data from our corpus and used them to annotate our corpus. Depending on the application and the specific

Table 3.1: Taggers trained on the Reuters corpus

Topic	$F_{0.5}$-Score	$F_{0.5}$ Std.Dev.	Precision	Recall	C
CRIME	78.92	1.51	82.93	66.59	1
DISASTERS	83.4	3.7	87.69	70.34	1
ELECTIONS	70.32	8.74	78.99	49.32	10
FASHION	83.88	18.61	94.61	71.27	0.01
PRICES	77.01	3.19	81.45	63.38	1
MARKETS	92.02	0.32	94.09	84.63	1
PETROLEUM	70.67	2.78	75.14	58.73	1
SCIENCE	73.63	5.17	83.72	50.62	0.01
SPORTS	97.78	0.5	98.31	95.75	0.1
WEATHER	71.43	3.68	82.91	46.84	0.01

problem we want to answer we can use a specific subset of these tags. For example for the demo in Sect. 3.3.2 we used classifiers trained on data from our corpus that work to a very high precision fashion, and for the experiments in Chapter 4 classifiers trained on the Reuters and The NY Times that treat Precision with double weight to Recall (with maximum $F_{0.5}$ score).

'Reuters' Taggers

We trained 10 taggers based on the Reuters dataset [116]. We used most of the corpus for training, namely from October 20st, 1996 until 1st July, 1997, plus the following six weeks for testing. For training we used up to 10K positive samples per class and equal number of negative samples randomly selected. Testing was on all data of the testing period. The cosine kernel was used for training and five different values, namely $0.01, 0.1, 1, 10, 100$, for the C parameter were tested. Table 3.1 presents the corresponding performances of the taggers after the β parameter was adjusted to achieve the maximum $F_{0.5}$ scores.

In Fig. 3.5 we use MDS and plot on a 2D plane, 100 randomly chosen articles for each one of the 10 topics in Reuters we track. We can observe the shape of topics in the space and their relative volumes. Furthermore we can observe overlaps between topics such as between Weather and Disasters

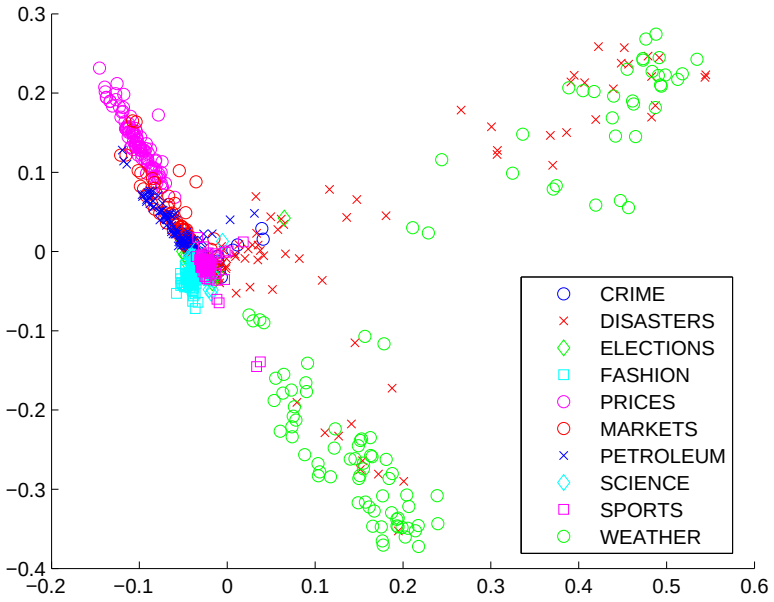


Figure 3.5: A visualisation of the ten Reuters topics we track. We applied MDS to project the articles on a 2D plane.

(*e.g.*, articles referring to disasters from hurricanes).

'The New York Times' Taggers

We used The New York Times corpus to train an additional set of taggers [157]. We used the last part of the corpus from 1st January, 2002 until 30th November, 2006 for training, and six monthly periods starting from December 1st, 2006 for testing. For training we used up to 10K positive

Table 3.2: Taggers trained on The NY Times corpus

Topic	$F_{0.5}$-Score	$F_{0.5}$ Std.Dev.	Precision	Recall	C
ART	81.67	1.34	84.9	71.38	1
BUSINESS	81.16	1.19	86.23	65.87	1
ENVIRONMENT	64.29	4.26	73.48	43.7	100
POLITICS	73.81	2.29	76.65	64.81	0.1
RELIGION	74.95	4.21	83.57	53.59	100

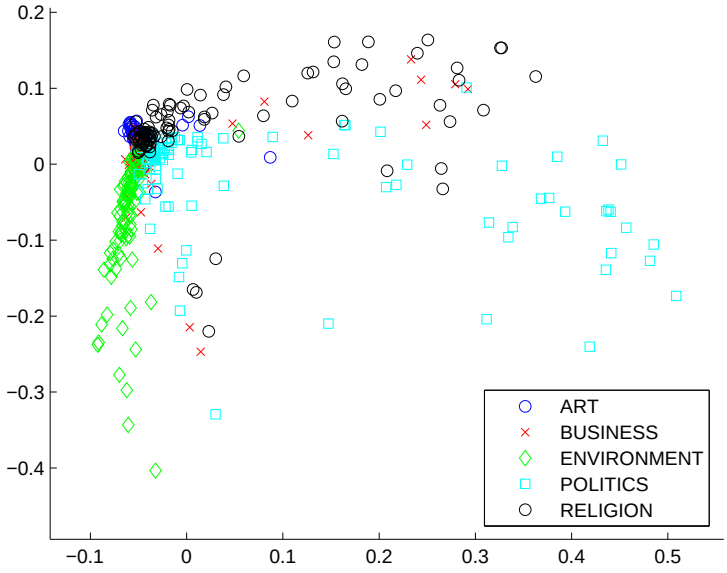


Figure 3.6: A visualisation of the five NY Times topics we track. We applied MDS to project the articles on a 2D plane.

samples per class and an equal number of negative samples were randomly selected. Testing was on all data of the testing period and for the same parameters as for the Reuters taggers. Table 3.2 reports the performances of these taggers after the β parameter was adjusted to achieve the maximum $F_{0.5}$ scores.

In Fig. 3.6 we use MDS to plot on a 2D plane, 100 randomly chosen articles for each one of the 5 topics in N.Y. Times we track.

High Precision Taggers

We also trained another set of taggers based on data from our corpus [65]. The annotation of the training set was based on feeds that are known to carry only specific topic news items, *e.g.*, politics. For these taggers we follow a philosophy of *positive tagging* meaning that a tag is given to an article if we are confident. The decision threshold of the classifier was set so as to maximize recall while achieving precision >99.0%. This is in order to

Table 3.3: Taggers trained on data form our corpus.

Tagger	Precision	Recall
Accidents & Disasters	99.74	90.65
General Science	99.32	42.23
Env. Issues	99.08	50.17
Business	99.33	38.19
Politics	99.65	55.87
Crime	99.90	54.57
War & Conflict	99.71	85.93
US-Elections	99.81	34.81

archive a low false positive rate, and apply the tag only to articles that it is most certain that they actually belong to the corresponding class. These taggers are retrained every month based on data of the previous month. For each tagger a balanced training set of up to 10K positive samples is used. Table 3.3 presents the performance of these taggers. In all cases it was possible to achieve a high precision of above 99%.

3.2.2 Keyword Detection

One interesting application of SVMs is their ability to discover the most important keywords that characterize a topic. To deploy the method we just need to evaluate the weight of each word in the vocabulary:

$$w = \sum_d \alpha_d \cdot x_d$$

where the summation is over each document x_d which is a Support Vector and belongs to the positive class. The vector w contains the weight of each word in vocabulary. The higher the weight the more relevant the word. Table 3.4 presents the top-20 most relevant stemmed words per topic for the taggers trained on Reuters and NY Times.

Table 3.4: Top-20 keywords per Reuters topics.

Topic	Keywords
CRIME	court, polic, alleg, legal, suit, lawsuit, case, investig, sue, litig, crime, arrest, file, settlement, corrupt, fraud, illeg, judg, prison, su, seiz, claim, fine, settl, appeal, jail, charg, scandal, rule, action
DISASTERS	kill, damag, crash, flood, drought, refineri, accid, caus, injur, blaze, earthquak, ship, disast, destroi, explos, typhoon, aground, di, eurotunnel, tremor, extinguish, emerg, unaffect, plane, hurrican, quak, blast, hit, death, storm
ELECTIONS	pct, labour, elect, pdi, vote, golkar, zealand, poll, jupp, clinton, parti, clark, candid, parliament, erbakan, emu, minist, blair, staf, auckland, coalit, gmt, britain, voter, programm, congratul, better, win, prime, polici
FASHION	fashion, design, benetton, versac, karan, colour, beograd, galliano, dior, givenchi, collect, catwalk, milan, armani, jacket, dress, model, donna, cloth, ferr
INFLAT.-PRICES	pct, inflat, price, cpi, statist, index, forecast, finnish, ppi, chang, gdp, wholesal, australia, rtr, consum, deflat, produc, deflationari, inflationari, econom
MARKET	tonn, pct, usda, goi, rate, bank, fix, lb, yield, dollar, rupe, auction, export, barrel, bours, bond, crude, newsroom, oil, trader
PETROLEUM	shr, energi, oil, plc, petroleum, crude, resourc, opec, russia, barrel, ga, explor, avg, rev, camco, shell, veba, ensco, halliburton, pipelin
RELIGION	church, islam, religi, ramadan, moslem, eid, priest, vatican, cathol, pilgrim, monsengwo, pope, teresa, mosqu, archbishop, pilgrimag, prai, cult, jewish, keng
SCIENCE	drug, trial, research, phase, patient, develop, space, studi, clinic, cancer, launch, scientist, mir, test, cell, clone, diseas, technolog, shuttl, gene
SPORTS	soccer, cricket, match, race, cup, rugbi, tenni, game, beat, championship, second, team, club, leagu, wicket, metr, win, titl, olymp, world
WEATHER	wsc, mph, weather, storm, rain, hurrican, cyclon, wind, tropic, temperatur, forecast, crop, ci, celsiu, harvest, frost, snow, flood, condit, ship

3.2.3 Prediction of Readers Preferences

In this section we present another area of interest to social scientists and media editors that can be affected by the application of Machine Learning techniques [87]. It is an approach to the question of what makes an article popular to a given audience, a question very close to the classical one of "What makes stories newsworthy?" [75]. This problem can be addressed today since the web based presentation of news allows the detection and capture of the exact preferences of audience simply by tracking the articles their clicks on a news webpage. The study of ranking users' preferences has been the area of research of data mining scientists and examples of related works include the study of popularity of stories in micro-blogging [174]; the study of users' attention to advertisements [76] and search engines log analysis [98].

We focus on 10 outlets that provide information about the most popular articles they provide in the form of a specialised 'Most popular' news feed. We compare what choices were available to the audience of an outlet basedd on the 'Main' feed of the outlet, and which articles managed to become popular. In this section we base our analysis only on the title and the description of each article, since only this information is available to the user when he decides to click or not to an news item. We tracked the 10 outlets from June 1st to December 31st, 2009 and gathered the articles of the two feeds of interest. For each outlet and for each day, we formed all possible pairs of articles that a) appeared on 'Main' feed only and b) appeared in both the 'Main' feed and the 'Most popular' feed, *i.e.*, pairs of popular and non-popular articles. In total we created 405,487 pairs, or 5942 pairs per outlet per month on average. For the'Florida Times-Union' we managed to successfully collect data for only five months and for the 'LA Times' for only three months.

The formed pairs of popular and non-popular stories were used as a dataset for a ranking task that was solved by Ranking SVM. It is worth mentioning that this problem can not be solved if treated as a binary classification problem, *i.e.*, by trying to separate the popular from the non-popular stories [88]. This is because popularity is a relative concept, that depends on

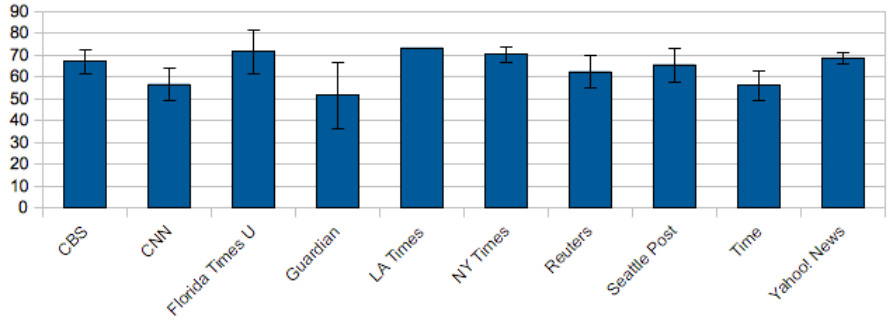


Figure 3.7: Average month-by-month prediction accuracies per outlet using Ranking SVMs.

what other stories are present on a given day to a given outlet, rather than an absolute one. The experimental framework we used is based on forming six training/testing sets per outlet, where each training set is one month and the corresponding testset is the pairs of the following month. We apply the SVM^{Rank} implementation of Ranking SVM [98] to predict the next months popular articles. We measured classification accuracy for three different values of C parameter, namely 1, 10 and 100, and we report the best achieved results per outlet in Fig. 3.7. We present accuracy and not precision/recall since the data are pairs and thus balanced by definition. The error bars are set to 95% confidence interval. We managed to get a significant performance for seven outlets, reaching up to 73.14% for 'LA Times'. The low performance for the 'Guardian' is explained by the limited amount of available data: only 482 pairs of articles on average per month rather than thousands as in the rest of the outlets. For two outlets, the 'CNN' and the 'Time', we managed to get only marginally significant results, and we assume the diversity of their audience and topic coverage to be the reason.

For the seven outlets for which we achieved a significant performance we can detect the most popular articles of each month. In Table 3.5 we present as an example the top three most popular articles for each outlet for December 2009. We can observe that each outlet has a different set of most popular stories and this can be attributed to a) the different choices of editors on which stories to cover and b) the differences on preferences of each audience

Table 3.5: Titles of most popular articles per outlet as ranked using Ranking SVMs for December 2009.

Outlet	Titles of Top-3 Articles
CBS	Sources: Elin Done with Tiger — Tiger Woods Slapped with Ticket for Crash — Tiger Woods: I let my Family Down
Florida Times-Union	Pizza delivery woman killed on Westside — A family's search for justice, 15 years later — Rants & Raves: Napolitano unqualified
LA Times	Pacquiao to fight Mayweather in March — Bone marrow transplant 'gets rid of' sickle cell anemia — Disney toys get Pixar animation guru's touch
NY Times	Poor Children Likelier to Get Antipsychotics — Surf s Up, Way Up, and Competitors Let Out a Big Mahalo — Grandma's Gifts Need Extra Reindeer
Reuters	Dubai says not responsible for Dubai World debt — Boeing Dreamliner touches down after first flight — Iran's Ahmadinejad mocks Obama, "TV series" nuke talks
Seattle Post	Hospital: Actress Brittany Murphy dies at age 32 — Actor Charlie Sheen arrested in Colorado — Charlie Sheen accused of using weapon in Aspen
Yahoo! News	Yemen is growing front in al-Qaida battle — Report: US helped Yemen's strike against al-Qaida — AP source: Al-Qaida operative killed by US missile

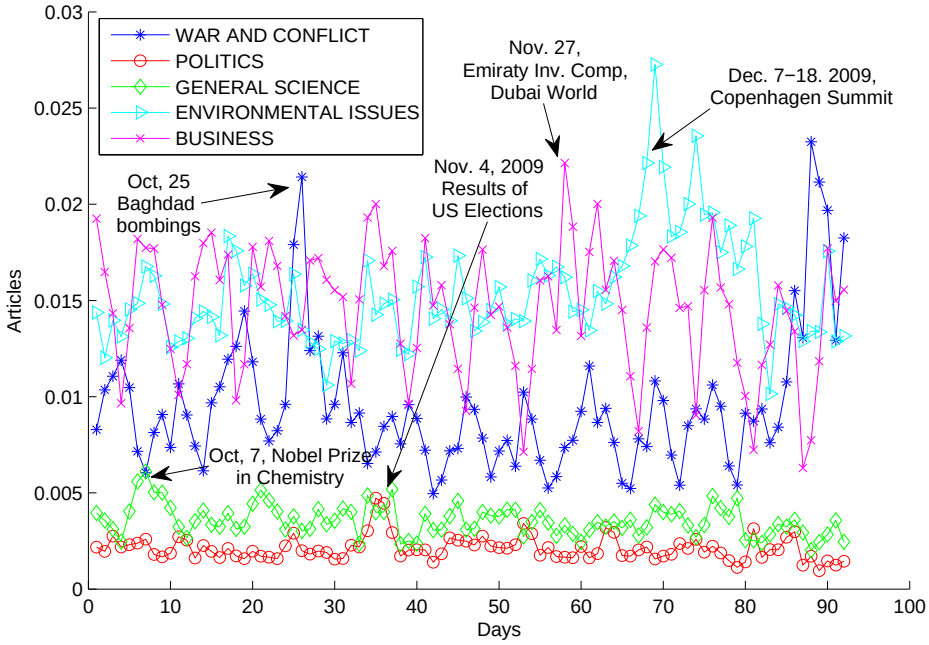


Figure 3.8: Three months monitoring of news. Major events relate to increase of volumes of tags of related topics. Article volumes are normalized over total number of news per day.

of the different media.

The performance we achieved is satisfactory given the limited available data and that we didn't took into account all other factors that affect readers' choices such as: the placement of the article in the layout of the web-page [190]; the presence of photos or videos; the presence of sentiment; or the novelty of the stories [174]. Even more, the segmentation of the audience is very coarse since the audience of each outlet is bundled into the same set and we focused only on the main page of each outlet – the page that covers stories that the editors expect to be interesting and popular for their audience.

3.2.4 Event Detection

A definition of an *event* can be "something (non-trivial) happening in a certain place at a certain time" [193] and has been a focus of the data mining

community. A precise approach to detect events can be based on the notion of *memes* [42, 17]. Memes, are defined as sequences of words that appear significantly more often than someone would expect by chance. Detection of memes in the news can be achieved using a suffix tree [80] as presented in [167, 165]. Other approaches include the detection and tracking of memes as found in quotations in the news [114]. Other methods for events detection include the detection of bursting terms [104]. This method has been recently applied for the detection of events in the blogosphere [148]. Others aim at tracking named entities or noun-phrases [173].

News topics, as general as those we track, can not be used for detection of specific events since they cover a much wider domain than this of a single event. Nevertheless, in this section, using an example, we want to show the potential of detecting major events in the news based on the measurement of the volume of topic tags. We tracked topics using the high precision taggers of the Table 3.3. Figure 3.8 illustrates the day-by-day monitoring of the last three months of 2009 and it presents the volumes of tagged articles, normalized over the total volume of articles for each day. It can be seen that a relevant increase in the number of articles tagged with a topic tag, relate to major event of that topic. For example a significant increase of science news can be related to the Nobel prizes awards, and a significant increase in political news can be related to the 2009 US elections.

3.3 Demos

We present two demos that were developed for demonstrating the ability of constantly monitoring the news media system, namely 'Found In Translation' and 'Science Watch'. The first has the goal to present multilingual news according to their topic on a 'heat-map' of Europe allowing an easy comparison of the topic bias in the media of the different countries of EU. The second to present scientific news categorised in six specialised topics. Other demonstration websites, built based on NOAM infrastructure include the Celebrity Watch [6] project which tracks celebrities; and the MemeWatch [167] for the tracking of memes.

3.3.1 Found In Translation

Found In Translation (FIT) is a web system that presents the multilingual news from the EU media translated into English and organised by topic and by country in a coherent manner [183][1]. For the translation of the news, Statistical Machine Translation techniques are utilised while for the categorisation of the translated articles into different topics Machine Learning techniques were used.

FIT also allows a comparison between the EU countries and the topics that the media of each country focus on. To enable a fair comparison, we only focus on the top-10 of the news outlets of each country. Their ranking is based on their Alexa rank score (See also section 2.1.1). We allow a sliding time window of the last seven days to track the articles published by the top-10 news outlets from each EU country; and we count their volume in each of the topics we track. We afterwards normalise these quantities by the total number of articles tracked in each country.

The normalised quantities are used to the web interface to colour the European map. A different heat-map for each topic, as illustrated in Fig. 3.9, is also generated. Each country presented on the map is coloured according to these normalised quantities: green colours are used to represent a 'weak' interest from each country's media for a specific topic, while red colours are used to indicate a 'strong' interest. FIT also provides a secondary visualisation, using bar charts, where the absolute numbers of the articles per topic and per country are presented.

Under the heat-map, FIT presents a list of the translated articles for each country and for each topic. We present the titles and the summaries of the articles while we also provide links to the original source of the article. We only use high translated quality articles by filtering the entries and including only those with up to one untranslated word. For UK and Ireland only those articles written in English language are shown. For multi-lingual countries like Belgium, the list of articles may contain news from more than one

[1]Found in Translation: http://foundintranslation.enm.bris.ac.uk (Accessed April, 1st 2011).

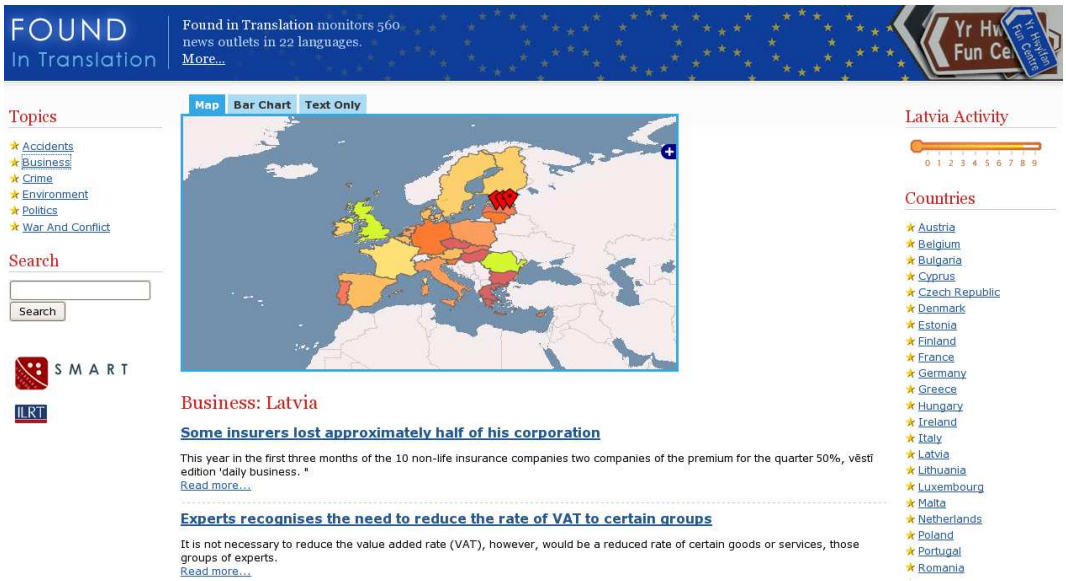


Figure 3.9: Found in Translation web interface.

different languages.

FIT's strength lies on providing access to an English speaker to the content of news media from other EU countries although written in different languages. It also gives the opportunity not only to read the news of other countries, but also compare the current interest of the media in each country for a specific topic. Comparisons like this provide an insight to the differences between the media of each country and their topic biases. The classification of multilingual news is a challenging task since it would require annotated training data in different languages. Our approach, of utilising SMT and then classifying, allows to focus only in one language, *i.e.*, English, and build the text classifiers for that language only where labelled data are easily available.

3.3.2 Science Watch

Science Watch is a weekly online science magazine entirely edited by a computer. It follows scientific news and mainly press releases issued by the press offices of universities, and report on research achievements at the institution. We follow UK, USA, and other English-language universities and research organisations, as well as public relations offices of funding bodies such as

Table 3.6: Science Watch Taggers performance

Tagger	Precision	Recall
Physics	99.90	91.66
Health	98.33	51.24
Space	99.67	48.25
Technology	99.88	63.72
Environment	99.68	29.47
Life Sciences	99.90	66.52

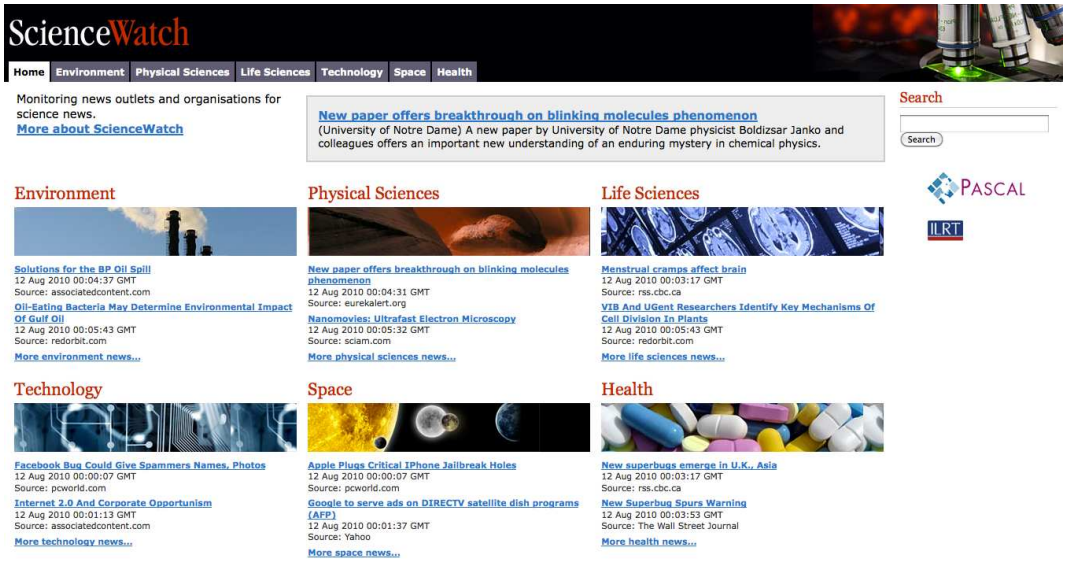


Figure 3.10: Science Watch web interface.

RCUK, NHS, NIH and NSF.

We collect news, automatically tag them, and organise them in six categories, namely Environment, Health sciences, Physical sciences, Life sciences, Space and Technology. These taggers are trained on data from our corpus in the high-precision/low-recall end of the spectrum. Table 3.6 presents the performance of the taggers currently used in Science Watch. Figure 3.10 illustrates the web interface of the demo. A next step of this project will be to analyse certain patterns in the content of scientific news items, per media and per topic, similarly to the work of media scholars [18].

3.4 Summary

In traditional media analysis annotation of news articles is done by hand, a laborious and time consuming work. In this chapter we showed that this process can be partially automated given a large set of pre-labelled documents that can be used as a training set for the classifier. We focused on detecting general news topics but the same approach can be used for the detection of any well defined topic given a large number of training examples.

The main advantage of using machines as coders is that they can be deployed on large scale, orders of magnitude higher than the reach of human coders. On the other hand human coders are more accurate and they don't require the large set of examples to learn. This annotation is only a first step towards answering questions about the media system at a higher level. The quality of the classifiers is not critical, since in a large scale study, if enough data are used, significant measurements can be achieved.

In this Chapter we also introduced Multidimensional Scaling that we are going to apply for several cases in our research, *e.g.*, for the detection of topic bias between outlets in Chapt. 4. In the next chapter we utilise the machine generated tags in order to answer questions about differences between different topics, *e.g.*, which is easier to read, politics or sports?

Chapter 4

Quantifying Properties of News Items

In this chapter we explore the application of Data Mining, Machine Learning and Natural Language Processing techniques for the automated analysis of biases in news media content [7]. These experiments are similar to studies performed by social scientists. The difference is that our approach is automated and in a scale out of reach of traditional methods. We based the experiments on a corpus of 2.5M news items, collected over ten months from 498 different English language news media from 99 different countries. We focus on detecting differences among topics in terms of their readability, their linguistic subjectivity and their popularity. Furthermore, as a case study, we focus on 16 leading newspapers and compare them based on their topic selection bias, their readability and their linguistic subjectivity.

4.1 Dataset

The dataset we use in these experiments is built using the NOAM system described in Sect. 2.4. We use a subset of news items that were published for a time period of ten months from January 1st, 2010 until October 30th, 2010, in the 'Main' feed of English language outlets. This resulted in 2,490,429 articles, from 498 English language news outlets, from 99 different countries.

Table 4.1: Articles per topic in the dataset for the period of the experiment.

Topic	Articles	Topic	Articles
ART	42896	MARKETS	24319
BUSINESS	126494	PETROLEUM	21236
CRIME	277626	POLITICS	201776
DISASTERS	83828	RELIGION	34441
ELECTIONS	28656	SCIENCE	10076
ENVIRONMENT	16103	SPORTS	141665
FASHION	1284	WEATHER	8505
INFLATION-PRICES	2331	Total	1037359

The 'coding' of such a large corpus of millions of articles is impractical to be performed by humans for reasons of cost and time. The automation of this process is highly desirable and simple tasks, such as topic annotation of news articles, can be automated by Machine Learning techniques. For this study we annotated the dataset based on taggers trained on Reuters and NY Times corpora (See Sect. 3.2.1). In total 1,037,359 topic tags were applied as illustrated in Table 4.1. Note that we allow each article to have more than one tags. The number of articles that remained untagged (concerning the Reuters trained taggers) are 1,564,018 and this is due to our choice to prefer high precision at the expense of recall. Furthermore, articles may belong to other topics than those that we track. Nevertheless, we have a total of 926,411 annotated articles, with an average of 1.12 tags each, which we can further analyse. The design decision of preferring a high precision at the expense of recall is justified because we want to represent each tag with as less noisy data as possible. The trade-off is a reduced number of tagged articles which is affordable due to the large volume of data we analyse and the fact that the tagged articles are more similar to what definition of each topic is used by the specific sources of the training data (Reuters and New York Times).

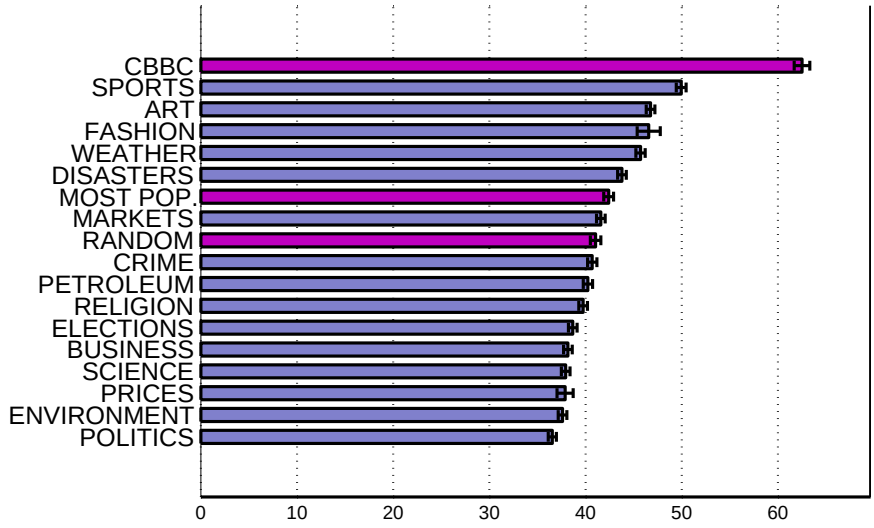


Figure 4.1: Readability of topics.

4.2 Comparison of Topics

We are interested in measuring and comparing properties of news items among different topics and outlets. We will focus on Readability, Linguistic Subjectivity and Popularity. The first two are stylistic properties of text. They are measured using NLP methods which are applied on a chunk of text and they result in a real number that measures the quantity. Popularity is based on the 'Most Popular' news feed as provided by some outlets.

4.2.1 Readability

Readability is the ease of comprehension of a text. The readability of news media has been in the focus of journalism scholars for a long time [41, 171, 100]. High readability is desirable by editors in order to increase readers satisfaction [72, 27].

We measured readability using the popular Flesch Reading Ease Test [67]. This test utilises linguistic properties of a text, namely the average sentence length (ASL) measured in words per sentence and the average number of syllables per word (ASW), and calculates the score F based on the formula

[67]:

$$F(x_i) = 206.835 - 1.015 \cdot ASL(x_i) - 84.6 \cdot ASW(x_i)$$

The constants of the formula are empirically tuned and depend on the language of the input texts x_i. For typical English texts, this score ranges from 0 to 100 and the higher it is, the more readable is the text. Indicatively, high scores of 90–100 correspond to texts easily understandable by an average 11-year-old student, scores 60–70 are easily understandable by 13-to-15 year old students and scores 0–30 are less readable and best understood by university graduates. The score has received some criticism for its lack of sophistication, nevertheless it is indicative of readability and serves our purpose of providing an example of a suitable framework for automated coding.

We measured readability on a sample of 10K articles per topic of interest, selected randomly. Figure 4.1 shows the mean readability score and error bars represent the standard error of the mean (SEM) with 99% confidence. We found that 'Sports' and 'Art' are the most readable topics. In contrary, 'Politics' and 'Environment' are the least readable. In the same graph we include a bar named 'Random' topic which is based on a random selection of articles. This in order to understand where the average readability lies. The 'Most Popular' bar is a random selection of popular articles. By comparing the two aforementioned bars we can see that mean value of readability of popular articles is higher than average.

For validation reasons we measured a set of articles from the BBC show CBBC-Newsround, which is a current affairs programme written specifically for children. As expected the CBBC news were found to be significantly the most readable set of articles compared to all topics with a mean readability score of 62.50 and SEM = 0.27.

4.2.2 Linguistic Subjectivity

The role of language in the way news is reported is a field of study of social scientists. As Fowler claims:

"News is a representation of the world in language; ...[language]

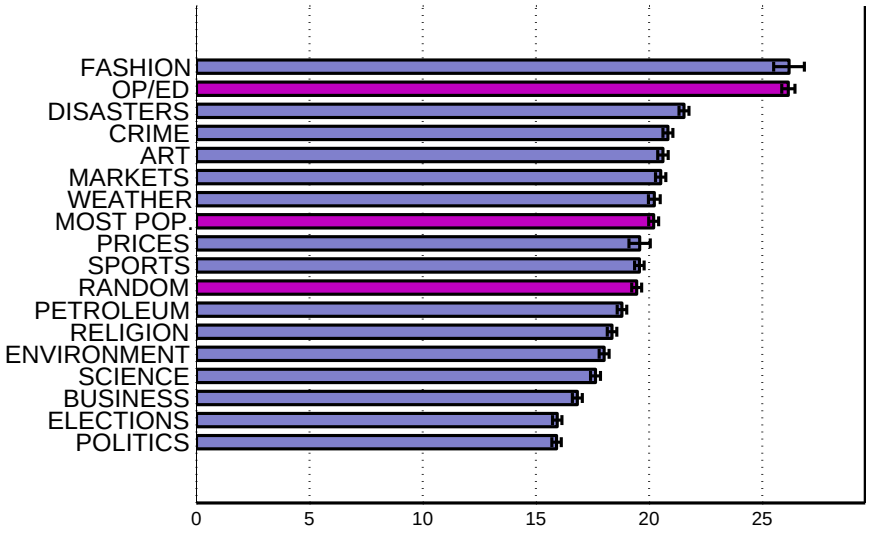


Figure 4.2: Linguistic Subjectivity of topics.

> imposes a structure of values on whatever is represented; [News] is not a value-free reflection of facts." [69]

Different words can be used to report the same news. The systematic choice of specific words over other alternatives can reflect some linguistic bias of the outlet. Natural Language Processing approaches can be used to measure that kind of biases.

We are interested in detecting the emotional value, or sentiment, associated with words present in news. A word can have a sentiment that is positive, negative, or neutral. Ideally a news article should be neutral, it should report only facts. In reality, news articles have content of polarised sentiment, either positive or negative. If this is the case we consider the article as *linguistically subjective*.

The metric of the extent of positive and negative sentiment of words has been numerically determined for common vocabularies of several languages [85]. Examples of computational research on the sentiment of news include: the separation of editorials that do carry subjective information, from business news, which is more factual [194]; study of the positive or negative way Named Entities appear in news [13, 77]; the use of sentiment in news to predict financial returns in markets [196].

In this study we measure the linguistic subjectivity by measuring the percentage of subjective adjectives over the total number of adjectives that are present in a news item. First, we detect the adjectives in the article using the "Stanford Log-linear Part-Of-Speech Tagger" [179, 180]. Then we check if these adjectives are indicators of sentiment, either positive or negative sentiment. We focused on the adjectives since they are associated with substantial sentiment bias [85, 86, 184]. To judge the level of linguistic subjectivity we used the latest SentiWordNet version 3.0 database, which is a list of words accompanied with two weights for sentiment orientation, varying from zero to one: one weight for the positive and one for the negative orientation [53, 10]. In this research we considered a word as subjective if either weight is above 0.25. If a word has many different values due to the different contexts it can be used, we calculated and used their average. We randomly selected 10K articles per topic and measured the linguistic subjectivity of their title and feed summary. We focus only on the title and description to measure linguistic subjectivity since the first paragraph is known to be based on the news angle and sets the tone for the rest of the article [54, 71].

Figure 4.2 illustrates the ranking of topics according to their linguistic subjectivity. We found that 'Fashion', 'Disasters' and 'Crime' articles are the most linguistically subjective, *i.e.*, most of their adjectives are indicators of sentiment. On the contrary, 'Politics' and 'Elections' are the topics with the least use of those adjectives. The 'Random' bar indicates a random selection of news articles, while the 'Most Popular' bar indicates the popular articles. It can be observed that popular news items are more linguistically subjective than average.

Proper validation of the approach is difficult in the absence of any agreed upon gold standard for sentiment analysis [144]. As an indicator of the validity of our approach we compared eight leading UK newspapers, four tabloids and four broadsheets, and verified the assumption that in general tabloids use more sentiment than broadsheets as described in Sect. 4.3).

Another evidence that corroborates our approach is based on comparing linguistic subjectivity of Editorial and Opinion types of articles to that of the

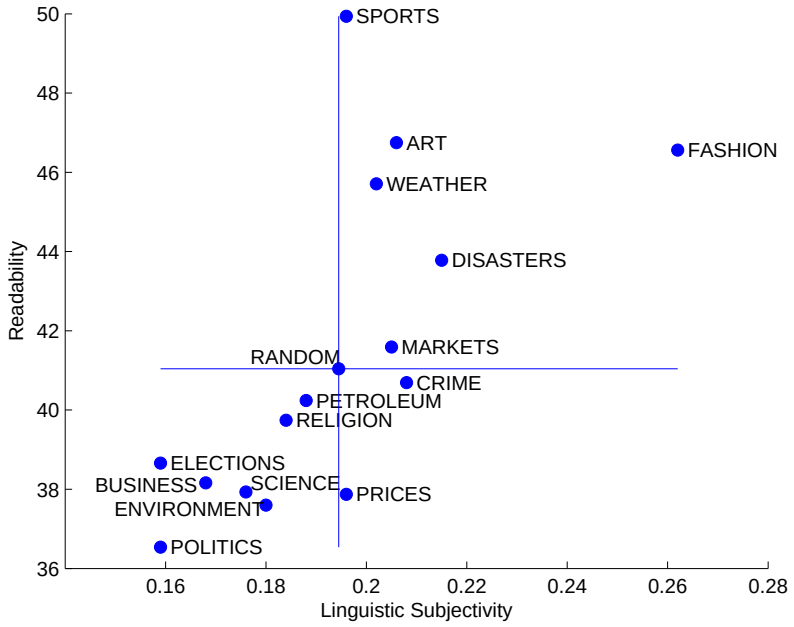


Figure 4.3: Scatter plot of readability vs. linguistic subjectivity of topics.

average news media content. We collected 5766 articles of that kind in our period of study from 57 different sources. We found that their linguistic subjectivity has a mean 26.15% with SEM of 0.29% while the mean subjectivity of main articles is 19.45% with SEM of 0.22%.

4.2.3 Readability vs. Linguistic Subjectivity

Readability and Linguistic Subjectivity are two properties that reflect the writing style of a text. In this section we explore their relation. In Fig. 4.3 we compare the readability of articles with their linguistic subjectivity for different topics.

We checked for correlation among those two writing style properties and we found a significant correlation of 72.5% (Spearman correlation, p=0.003). That is, the most readable topics tend to be also the most linguistically subjective and vice versa.

It is worth mentioning that we found that popular articles are more read-

Table 4.2: Outlets with Most Popular feed that we used for the analysis of topics popularity.

Outlet	Main Articles	Popular Articles	Main and Popular
Reuters	7100	5236	1775
Seattle Times	28437	3379	665
CBS	10665	9432	1933
The New York Times	21745	2829	898
CNN	8055	6550	289
Florida Times-Union	1212	6555	370
Seattle Post	10145	12711	504
Los Angeles Times	6696	1301	406
Forbes	13371	1836	1091
Time	5514	5793	2757
KSBW	4481	5086	3185
Yahoo!	18883	23101	7490
Guardian	15393	2608	58
News.com.au	4974	1282	499
The Wall Street Journal	12300	1185	370
BBC	31779	4072	2119
Total	200750	92956	24409

able than average news and more linguistically subjective than average news. This means also, that both these properties could be exploited as features for the identification of popular articles along the lines presented in Sect. 3.2.3.

4.2.4 Popularity

Our goal is to measure the conditional probability of an article to become popular given its topic ($\Pr(Popular \mid Topic)$). We calculate this probability for each one of the topics we track. It is calculated by the equation:

$$\Pr(Popular \mid Topic) = \frac{\Pr(Popular \cap Topic)}{\Pr(Topic)}$$

where the numerator is the probability of an article that appeared in the main feed of an outlet to be tagged as popular and also tagged with a given

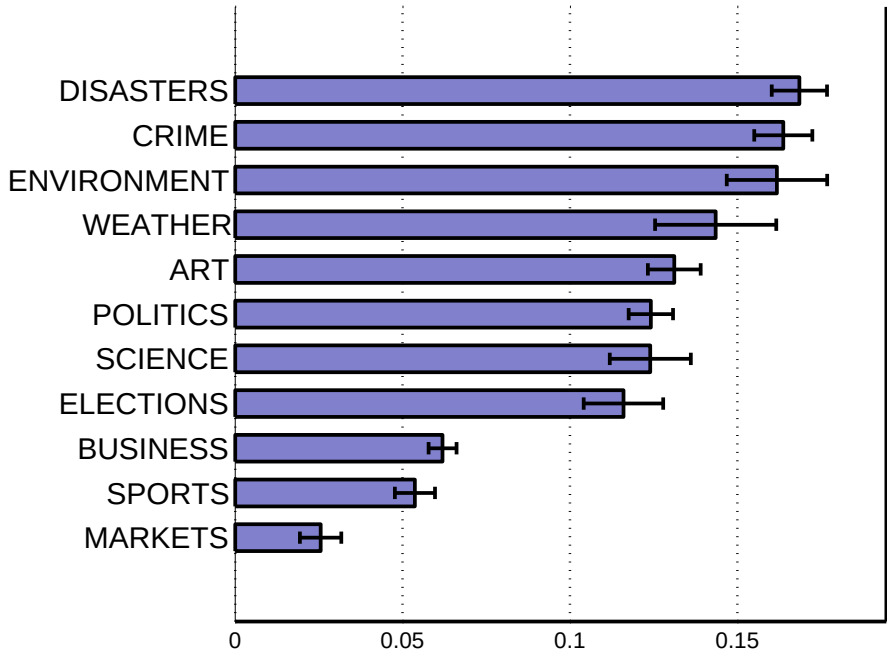


Figure 4.4: Popularity of topics. Errorbars are ±1 standard deviation of the mean. Topics for which we had not enough data to conclude with adequate confidence were removed.

topic. The denominator is the probability of an article that appeared in the main feed of an outlet to be tagged by the given topic.

To get popularity of articles we tracked the special feed that is provided by some outlets that carries the 'Most Popular' articles. Among the outlets we track, only 16 provide this special feed for the period of study. We collected a total of 92,956 popular articles, 200,750 articles that appeared in the Main feed of those outlets and 24,409 of those appeared also in the main feed of the outlet. The number of articles per outlet is shown in Table 4.2.

Figure 4.4 presents the ranking of news topics based on their popularity. We removed the topics for which we had not enough data to conclude with adequate confidence. The most popular topics, among those that appear in the main pages of outlets, are the 'Disasters' and 'Crime', while the least popular appear to be 'Sports' and 'Markets'. Of course these results refer to measurements only on the 16 outlets we tracked.

Table 4.3: The 16 newspapers of our case study and the corresponding number of articles that appeared in their main pages for the period of study.

Newspaper	Articles	Newspaper	Articles
Chicago Tribune	5477	Daily Mail	24326
Daily News	2212	Daily Mirror	7731
Los Angeles Times	6696	Daily Star	8946
New York Post	32033	Daily Telegraph	22682
NY Times	11508	Independent	43557
The Wall Street Journal	12300	The Guardian	15393
The Washington Post	7228	The Sun	9048
USA Today	6208	The Times	2957

4.3 Comparison of News Outlets

The last experiment we conduct in this chapter is a case study on a subset of 16 outlets, eight UK newspapers and eight US newspapers. We measure their topic selection bias, and their writing style preferences in terms of readability and linguistic subjectivity. We intend to show that our automated approach can be applied also for the direct comparison of outlets. For the 'Daily Star' we take into account the main news feed which is separate from the 'Celebrities' feed.

We attempt a comparison of US and UK newspapers, as well as between UK tabloids and UK broadsheets. Differences of tabloids to broadsheets are well studied by media scholars and include differences in range, form and style [133, 185], changes in journalistic behaviour [52], *etc*. Broadsheets tend to use emotionally controlled language [71]. It has been argued that nowadays differences between tabloids and broadsheets are smaller with broadsheets starting to adopt styles of tabloids [70].

4.3.1 Topic Selection Bias

We can represent each outlet as a vector of 15 dimensions, one for each topic we track. The value of each dimension is the number of articles of the outlet tagged for the given topic and normalised over the total number

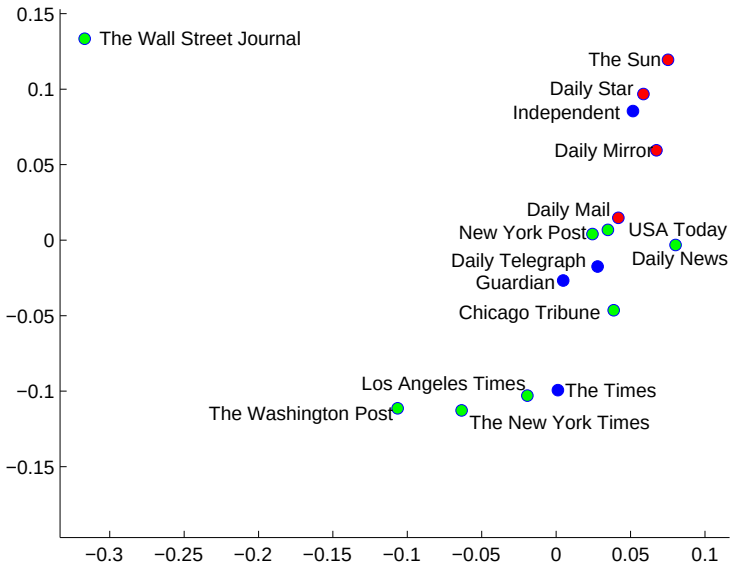


Figure 4.5: Topic selection bias of 16 leading newspapers. US newspapers are coloured in green, UK broadsheets in blue, and UK tabloids in red. Outlets that tend to select the similar topics in their front pages are located closer to each other.

of articles in the outlet. We can measure the euclidean distance of these vectors and use Multidimensional Scaling (MDS) to plot them on a 2D plane. This visualisation, illustrated in Fig. 4.5, reveals the similarity of the outlets on their topic selection bias: outlets that cover the same topics are closer together, while outlets with different topic selection criteria are further apart. One easy observation is that UK tabloids cluster together in the plot since they tend to cover the same topics in their frontpages.

One interesting outlier in Fig. 4.5 is the *The Wall Street Journal* which is significantly apart from all other outlets. We can assume that this due to the strong bias of that outlet in favour of business news. To validate this hypothesis we checked what the MDS output would be, if we first removed all articles in dataset tagged with the 'Business' tag. This plot is illustrated in Fig. 4.6. We can observe that there are no more outliers and this is a strong indication for the validity of our assumption.

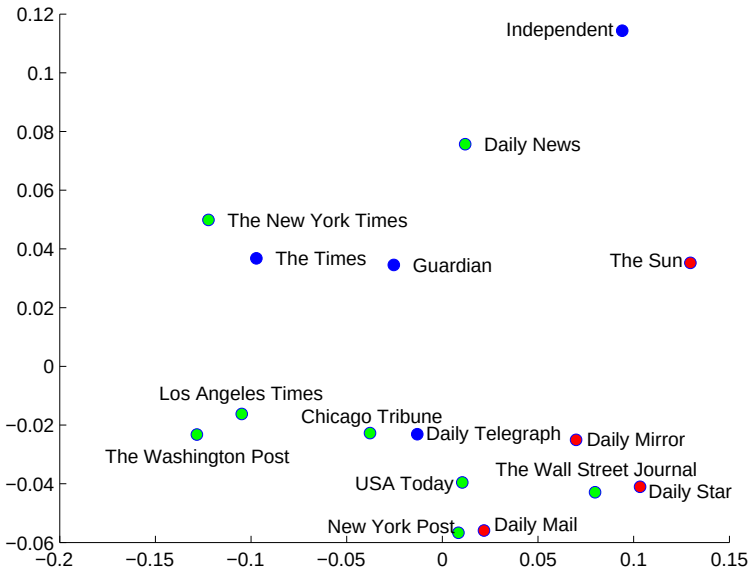


Figure 4.6: Topic selection bias of the 16 newspapers if we ignore and remove from the dataset the articles tagged with Business.

The afore used method for the visualisation of the topic selection bias of outlets can easily scale up. In Fig. 4.7 we illustrate the topic selection bias for all 498 outlets in dataset.

4.3.2 Readability and Linguistic Subjectivity

Figure 4.8 presents the readability of the newspapers and Fig. 4.9 their linguistic subjectivity. Error bars represent the error of the mean and are practically zero for linguistic subjectivity. We found that the most readable newspapers are the UK tabloids *The Sun* and *Daily Mirror* while the most difficult to read appear to be *The Guardian* and *USA Today*. The same two tabloids use many adjectives with some sentiment orientation while *The Wall Street Journal* and *The Washington Post* use the least. In general, tabloids in UK use more sentiment in their articles compared to broadsheets.

Furthermore we created a scatter plot for the two writing style properties illustrated in Fig. 4.10. This can be seen as plot that reflects the writing style

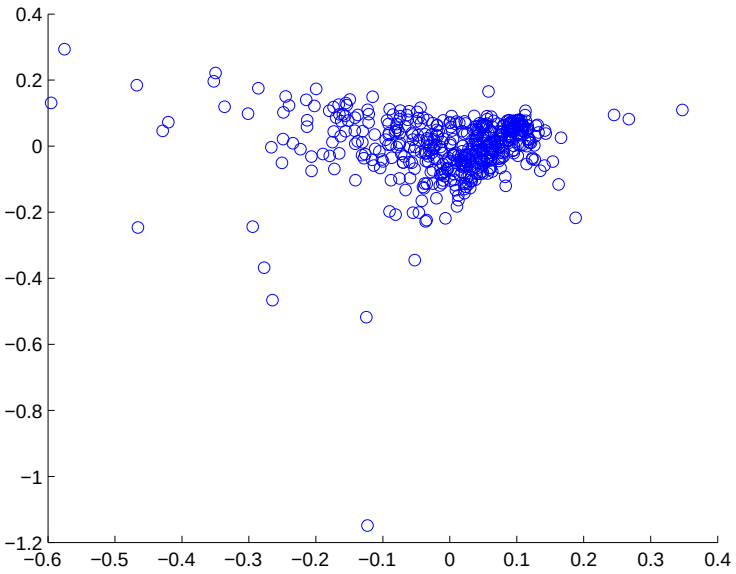


Figure 4.7: Topic selection bias of all outlets in dataset. We project the choice of topics that media choose to cover in a 2D plane. Each point on of the plot is an outlet and outlets that make similar choices of the news topics that they cover are plotted close together.

of each outlet. Outlets that are close have similar writing style in terms of readability and linguistic subjectivity. That is they target the same audience – audience with the same taste in news. Indeed several small clusters of newspapers can be identified: *The Sun* and *Daily Mirror*; the *The Times* and *N.Y.Times*; *Independent* and *New York Post*. There also seems to be an absence of high readability, low linguistic subjectivity news.

4.4 Summary

Our findings are the result of a completely automated process where human interaction is very limited. This study is intended to illustrate how different modern AI technology can be deployed to automate such diverse tasks, in order to enable large scale studies that would otherwise be impossible.

The key findings in this chapter based on our analysis are:

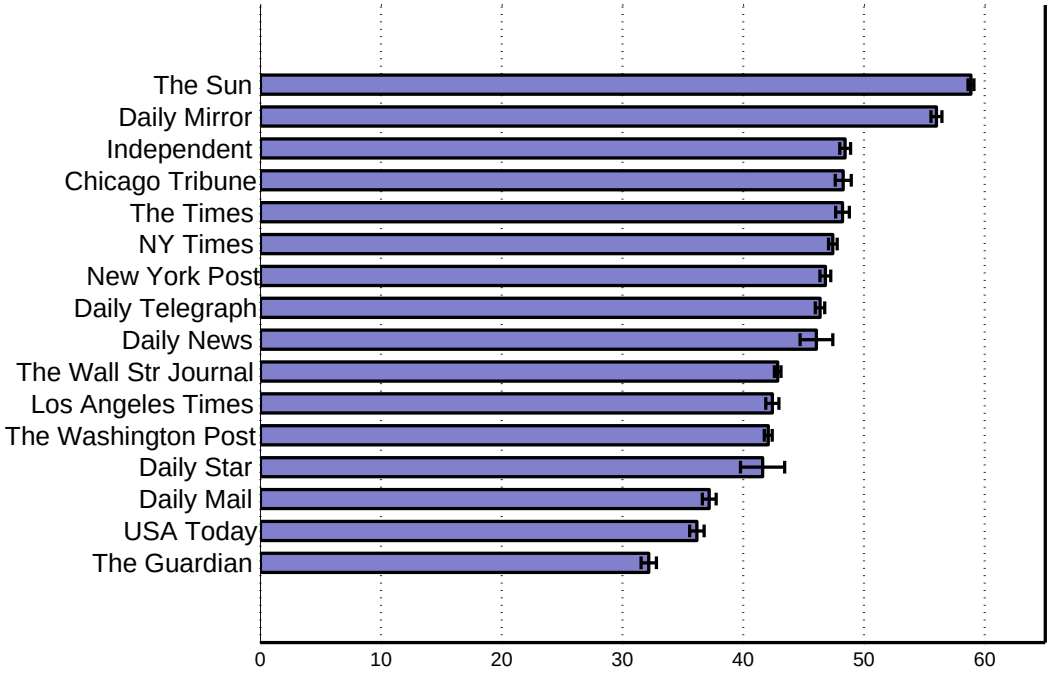


Figure 4.8: A subset of leading US and UK newspapers ordered by their readability.

- The most readable topics are 'Sports' and 'Art', while the least readable are 'Environment' and 'Politics'.
- The most linguistically subjective topics are 'Fashion' and 'Disasters', while the least linguistically subjective topics are 'Politics' and 'Elections'.
- There is a significant correlation of 72.5% (p=0.003) between the two styling properties of readability and linguistic subjectivity as it is measured between different topics.
- The most popular articles appear to be 'Disasters' and 'Crimes', and the least popular appear to be 'Markets' and 'Sports'.
- The comparison of leading UK and US newspapers based on their topic selection bias. UK tabloids tend to cover similar topics compared to other newspapers in the study.

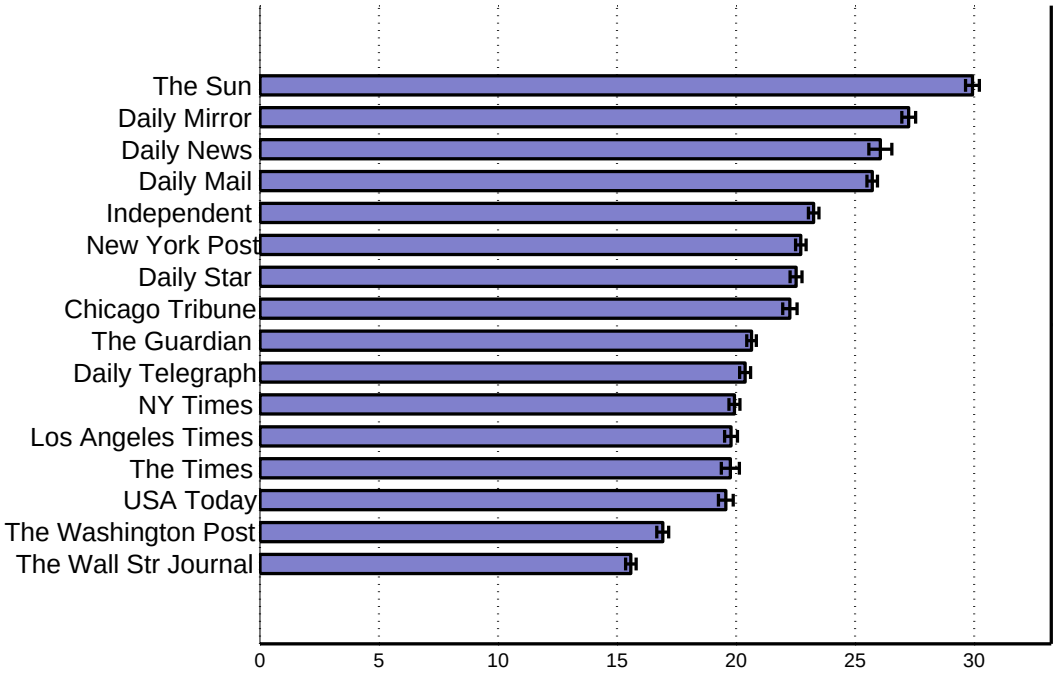


Figure 4.9: A subset of leading US and UK newspapers ordered by their linguistic subjectivity.

- The comparison of the newspapers based on their writing style in terms of readability and linguistic subjectivity.

Future work includes the measurement of the 'modality' of news. This is the use of modal verbs such as *may, could, should, will, must* and their negations, that reveal the degree of confidence of the writer to the claims they make [71, 69].

While this study perhaps points in the direction of the automation of media content analysis, we should also discuss the limitations of this kind of approach. Firstly, it is currently limited to 'shallow semantics', which offer more information than plain string-matching methods, but may miss nuances in the meaning of articles. Also, it is perhaps less accurate than human coders. On the other hand, this approach compensates for these drawbacks by enabling researchers to access vast datasets, and hence to apply the law of large numbers and statistical error correction. Furthermore, even if it is less accurate than human coding, certain experiments would not be possible

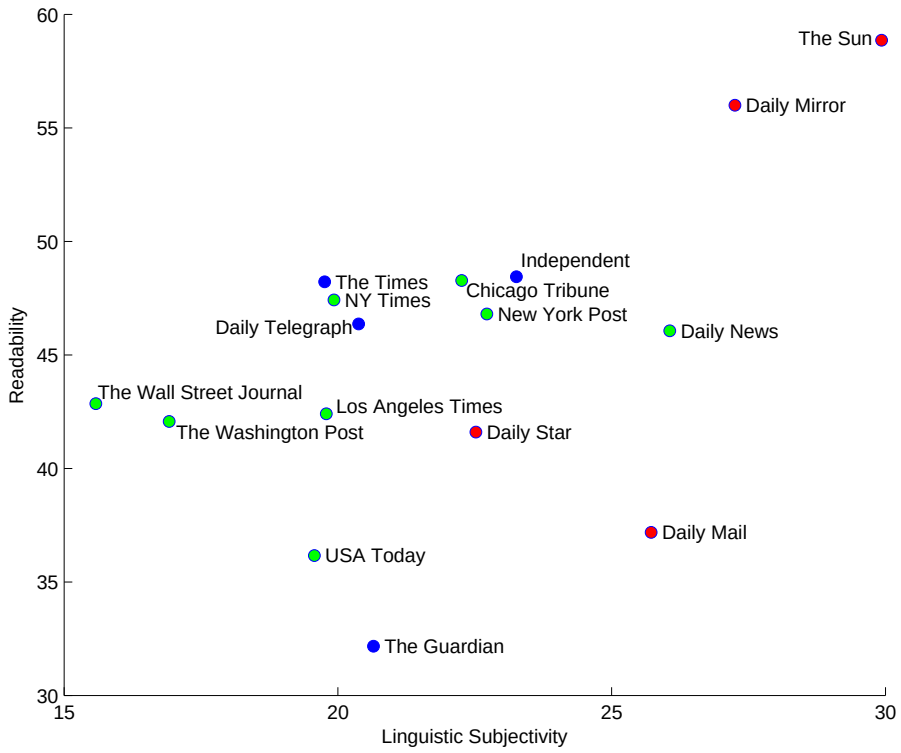


Figure 4.10: Readability vs. linguistic subjectivity of selected news media. UK tabloids are coloured in red, UK broadsheets in blue, and US newspapers in green.

if they were not automated.

Chapter 5

Network Inference from Media Content

In this chapter, we present the inference and the validation of the network of news outlets. This network is comprised of outlets as nodes and two outlets are connected if they publish similar stories [66]. We develop a general statistical framework to compare different network inference algorithms and validate the resulting network. The validation is based on statistical significance tests. We argue that the inferred network should be stable; it should be related to any available reference networks that we expect to be partially relevant; and that it should have a predictable structure. We compare three network inference methods for the network of news outlets and we show that all three validation properties stand true for one of those methods.

5.1 Network Inference and Validation

Network inference is an ubiquitous problem. It appears in many different fields ranging from physics and biology to social sciences. Examples of networks from natural sciences include gene regulatory networks [45]; biochemical networks [124]; and protein interaction networks [95, 145]. In social sciences we find examples of networks such as the Instant Messaging network [115]; Facebook friendships [28]; and the blogosphere [1]. The nodes

of these networks are genes, proteins, people or blogs. The edges in these networks represent some relation between the connected nodes, for example co-expression of genes; binding of proteins; social contact of people; or similar content in blogs.

In all those networks we can observe the state of each node, but the overall underlying topology of the network is hidden. Several methods and approaches have been proposed regarding reconstructing, or inferring, these networks depending on the domain and the available information in each case. If more than one approach for inferring the network is available and each one produces a different result, and if there is no ground truth to compare to, then the selection of the most appropriate method becomes really challenging.

The output of a network inference method can be validated by similar approaches as the validation of any pattern found in data. Validation of patterns can be achieved by either assessing the significance of the pattern or by measuring their predictive power. The significance of a pattern depends on the probability that such a pattern could be found in randomly generated data. The predictive power of a pattern depends on the extent to which patterns found in a subset of the data can be found in another independent subset of the data. The two approaches are highly related but for our purposes we will address them separately. We refer to them as the 'Hypothesis Testing' and the 'Predictive Power' approach respectively.

5.1.1 Hypothesis Testing

We argue that the inferred network which is the product of a network reconstruction algorithm needs to satisfy two key properties. First, the inferred network needs to be stable, meaning that networks inferred from independent and identically distributed data must be similar to each other. For the case of the news outlets network we will show in the following sections that network reconstruction algorithms can produce similar networks when independent and identically distributed data are used for their production. Second, the inferred network needs to be related to any available ground truth for which is known or assumed it affects the network topology. For

the case of the news outlets network we will show that the inferred network is significantly related to some directly observable networks of news outlets constructed using information such as the country of origin or the language of each outlet.

Both of these two properties can be verified by using the same methodology, *i.e.*, by testing the hypothesis that the inferred network is similar to a reference network. Hypothesis testing is based on the key idea of measuring the probability that some observation in the data can also be found in random data. To perform hypothesis testing we are based on the notion of statistical significance, as expressed by the p-value. We need to define a test statistic that quantifies how similar is the inferred network to the reference network; and a null model for the generation of random networks similar to the inferred network.

We adopt as test statistic $t_{\mathcal{G}_R}$ the similarity of the inferred network to a chosen reference network $\mathcal{G}_R$. The null hypothesis, denoted by H_0, is that the inferred network is sampled from some underlying distribution. We then quantify the probability that a random network is at least as similar to a chosen reference network as the inferred network:

$$p \;=\; P_{\mathcal{G}\sim H_0}(t_{\mathcal{G}_R}(\mathcal{G}) \geq t_{\mathcal{G}_R}(\mathcal{G}_I)) \tag{5.1}$$

where $\mathcal{G}_I$ is the network inferred by the inference algorithm. A small p-value indicates that we can reject the null hypothesis and accept the alternative hypothesis. This means that the inferred network is more similar to the reference network than expected by chance. More formally, we make an arbitrary choice of a significance threshold α that depends on the application and we reject the null hypothesis if $p < \alpha$. In the above formula of p-value calculation we used a measure of distance as test statistic. If we use a measure of similarity instead of a distance, then the inequality signs of the equation should be inverted.

In practice it is difficult to compute the p-value exactly. However, we can estimate its value by sampling a large number of K networks from the H_0. Then the p-value is measured as the fraction of those random networks for

which the test statistic is smaller than that for the inferred network:

$$p \approx \frac{\#\{\mathcal{G} : t_{\mathcal{G}_R}(\mathcal{G}) \geq t_{\mathcal{G}_R}(\mathcal{G}_I)\} + 1}{K + 2} \tag{5.2}$$

The +1 term in numerator and the +2 term in denominator are due to the application of the Laplace correction to the calculation of the p-value. The Laplace correction provides a more robust calculation of a probability for small datasets by assuming that each possible outcome will occur at least one time. For example, for the calculation of a probability of an event e with $p_e = \frac{k}{N}$, the application of the Laplace correction will result in $p_e = \frac{k+1}{N+\|e\|}$ where $\|e\|$ is the cardinality of the different possible events. The Laplace correction is applied in several machine learning techniques such as in the case of supervised classification by decision trees [152].

In the next three sections we will discuss the issues of selecting a test statistic, the null model and the reference networks.

5.1.2 Test Statistics

As a test statistic we adopt the proximity of the inferred network to a reference network. The proximity can be quantified by using some measure of similarity, or distance, between networks.

Several measures of networks similarity have been proposed in literature[147, 25, 60]. The comparison of two networks $\mathcal{G}_A$ and $\mathcal{G}_B$ comprised of the same set of labelled nodes $\mathcal{N}$, can be performed by comparing their link structure. One of the simplest approaches is to directly count how many pairs of the same corresponding nodes between the two networks have the same linkage status, *i.e.*, they are connected or disconnected. This way the comparison if the networks is reduced to a comparison of the sets of edges of the two networks. A measure of dissimilarity between sets is their Jaccard distance [93]:

$$JD(\mathcal{E}_A, \mathcal{E}_B) = \frac{|\mathcal{E}_A \cup \mathcal{E}_B| - |\mathcal{E}_A \cap \mathcal{E}_B|}{|\mathcal{E}_A \cup \mathcal{E}_B|} \tag{5.3}$$

where $\mathcal{E}_G$ is the set of edges of network G. The Jaccard distance ranges between zero and one. A distance of zero indicates two identical sets and a

distance of one indicates two sets without any common elements.

Other more sophisticated measures of network similarity can be defined. For example, the comparison of two networks can be based on the comparison of the number of their triplets with the same connectivity status, *i.e.*, by comparing their network motifs [138]. For our experiments we adopted as a test statistic the Jaccard Distance of the inferred network to a reference network.

5.1.3 Null Models

A statistical test measures the probability that a pattern is significant and not the result of pure chance, generated by some random process. To measure this probability we have to formally define the notion of 'chance'. This is achieved by defining a null model that provides a baseline for the assessment of the patterns under study. Depending on the application one or more null models can be defined. In our study we need to specify a null model for the generation of random networks. The goal is to measure how close is the inferred network to a reference network, and compare that similarity to the similarity of the inferred network to some random networks. If the two levels of similarity are close then we can not conclude that the inferred network is similar to the reference network, since that could be just by pure chance. Next, we will introduce two methods for generating random networks, one basic approach and one more realistic one.

The first method we discuss is the celebrated $G(n, p)$ Erdős-Rényi model [51]. According to this model a graph of n nodes is created by connecting pairs of nodes randomly, with every two nodes having an equal probability of p to be connected. The higher is the probability p, the more dense graph is created. The expected number of edges e in that graph is:

$$e = \binom{n}{2} p \tag{5.4}$$

and the degree of any particular node A follows a binomial distribution:

$$P(deg(A) = l) = \binom{n-1}{l} p^k (1-p)^{n-1-l} \tag{5.5}$$

The Erdős-Rényi model is simple to analyse, but it leads to rather unrealistic networks that often have very different topologies than those observed in real world networks. For example, it does not create power-law degree distributions that are found in social networks [55].

A more realistic model would preserve some topological properties of the network under study. Such a desired property is the degree distribution of the nodes. A model that preserves degrees of nodes by construction is based on the Switching algorithm [139, 137]. The algorithm starts from a given graph and produces a randomized version of it by switching edges between nodes following the simple rule: if edges $A \to B$ and $C \to D$ are present, replace them with the edges $A \to D$ and $C \to B$. This process is repeated multiple times until a random network is created. A safe number of iterations is considered 100 times the number of edges of the network [137]. Similar randomisation can be deployed for undirected graphs.

5.1.4 Reference Networks

In our framework a key point is the selection of the reference networks. We utilise reference networks, since the real ground truth network is unknown. If we know the ground truth we can directly compare to that network. If it is unknown we can utilise other networks, or sub-networks, which are known, and which we expect to be related to the network we try to infer and validate. In the case of the news media network we will utilise as reference networks, networks constructed with outlets as nodes and simple rules for their connectivity. For example the 'Language' network is constructed by connecting two outlets if they share the same language. In the same way we will build the 'Location' network where two outlets are connected if they are from the same country and the 'Media-Type' network where two outlets are connected if they are of the same type (e.g. they are both newspapers). These

networks are comprised of cliques, one clique for each country for the case of the 'Location' network or one clique for each language for the 'Language' network.

5.1.5 Structure Prediction

In this section we will present a different approach for validating an inferred network. We investigate whether it is possible to predict the network topology based on information provided by the reference networks. We expect that those reference networks, that are accepted as ground truth, are able to predict some of the inferred network topology and 'explain' the existence or the absence of some network edges. If more than one reference networks are available we will present how this knowledge can be combined using Generalized Linear Models (GLMs) and predict the inferred network topology.

The GLMs were introduced by J. Nelder and R. Wedderburn as a unified framework for various non-linear or non-normal linear variations of regression [141]. Under GLM framework the model for the observed data Y_i is split into a random and a systematic component using a function called the 'link' function g. The observed data is assumed to be generated from a distribution function of the exponential family [131]. The observed data and the independent variables $\mathbf{X}$, are connected through:

$$E(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\eta) = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) \tag{5.6}$$

were $E(\mathbf{Y})$ is the expected value of $\mathbf{Y}$; $\boldsymbol{\mu}$ is the mean of the distribution; and η is the linear predictor which is a linear combination of unknown parameters $\boldsymbol{\beta}$. The elements of $\mathbf{X}$ are typically measured experimentally and in our case depend on the reference networks or any other useful information that is available for the inferred network. The unknown parameters $\boldsymbol{\beta}$ can be estimated by maximum likelihood or other techniques.

We measure the ability of a GLM model to predict whether two nodes of the network are connected. We treat the GLM as a classifier and assess its performance using a methodology similar to this followed in supervised classification approaches. The network nodes are split into two sets, a training

set and a test set. The training set of nodes and the edges that connect them are used to tune the GLM parameters in a training first. Then in the test phase the GLM is assessed on its ability to predict whether nodes in the test set are connected or not. Using this approach the edges that connect nodes among the training and the test set are not used, neither in training nor in testing phase. We make the assumption that this doesn't introduce any significant bias: it is similar as ignoring a small random set of the available training samples in a typical classification task. Furthermore the separation in training and test sets is repeated multiple times as in a cross validation scheme in order to reduce any potential bias on the estimations. The performance of a GLM model is assessed using the Area Under Curve (AUC) [57, 58, 61].

5.2 The Mediasphere Network

In this section we will illustrate the application of our network validation methodology for the task of inferring the network of news outlets that are connected by if they tend to cover the same stories. We will present three simple network reconstruction algorithms and we will apply our framework in order to select the method that creates the most stable network in time; that is most related to the available ground truth; and we can predict its structure. Similar approach can be used in other domains for example for the inference of gene regulation networks.

For the experiments in this section we analysed a subset of 1,017,348 articles gathered over a period of 12 consecutive weeks from October 1st, 2008, from 543 online news outlets, distributed across 32 different countries, in 22 different languages, including the different media types such as newspapers, blogs, *etc*. We clustered the articles using the Best Reciprocal Hit (BRH) clustering method, as described in Sect. 2.5. The clustering process detected 81,816 stories in total, an average of 974 stories per day for the period of study.

5.2.1 Network Reconstruction algorithms

In this section we will describe three simple network inference algorithms for the new media outlets network [66]. All algorithms use the same set of outlets as nodes, and in principle they try to connect pairs of outlets that cover the same stories. The algorithms differ in the specific policy they use to decide if two nodes should be connected or not. Each algorithm outputs a real value for each pair of nodes. Then a threshold has to be chosen for the desired density of the resulting graph.

Method A

A very basic method for the network reconstruction is to simply count for each pair of outlets how many stories they both publish. We then connect them if they share a minimum amount of common stories. This method doesn't take into account the total number of stories that an outlet publishes. Also it can easily lead to very dense networks even with a small threshold since due to the power-law we discussed in Chapter 2, there are few stories that are covered by the majority of the media.

Method B

A more sophisticated would take into account the popularity of a story among the different outlets. This approach is inspired by the TF-IDF method: each outlet correspond to a document, and each story corresponds to a word. Then, the frequency of story j that belong to outlet k is

$$f_j^k = \frac{s_j^k}{s^k} \tag{5.7}$$

where s_j^k is the number of times the story j appears to outlet k and s^k is the total number of stories of outlet k. The corresponding inverse outlet frequency i_j^k is defined as

$$i_j^k = \log \frac{n}{n_j} \tag{5.8}$$

where n is the total number of outlets and n_j is the number of outlets that carry story j. For each outlet we can construct a vector w^k, of size J equal to the total number of different stories, with one weight for each story:

$$w_j^k = f_j^k \cdot i_j^k, \quad j = 1, 2, \ldots, J \tag{5.9}$$

Then the similarity of two outlets N_a and N_b can be measured using their dot product:

$$sim(N_a, N_b) = \sum_{t=1}^{J} w_t^a \cdot w_t^b \tag{5.10}$$

Method C

A different approach is to assign a weight f_j to each story based on its popularity among different media:

$$f_j = \frac{1}{n_j} \tag{5.11}$$

where n_j is the number of outlets that have story j. The f_j assigns a small weight to stories that are covered by many media, and larger weight to stories covered by few media. The maximum weight of a story is 1/2 since by definition we consider as story, news that is carried by at least two different media outlets. The minimum weight is $1/n$ where n is the total number of outlets. Using the minimum and the maximum values we can scale the f_j weights to a range from zero to one:

$$f_j' = \frac{2(n - n_j)}{(n-2) \cdot n_j} \tag{5.12}$$

Using the scaled weights for each story we can define a measure of similarity between outlets:

$$sim'(N_a, N_b) = \frac{\sum_{t=1}^{J} f_t' y_a(j) y_b(j)}{\sum_{t=1}^{J} y_a(t) y_b(t)} \tag{5.13}$$

where $y_k(j)$ is an indicator function with value one if outlet k has story j and zero otherwise.

5.2.2 Validation

We compare the three algorithms described in Sect. 5.2.1 of networks reconstruction in terms of the stability of their outputs in time, how their results compares to related networks, and how well can they predict the network structure.

Stability in Time

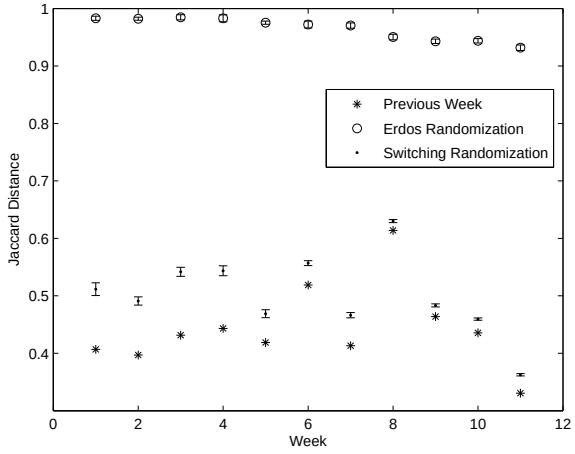
To test the stability of the network reconstruction algorithms in time we split the 12 weeks datasets in 12 sets, one for each week. We then create the inferred network for each weekly set using the three network reconstructions methods and we set as desired density ~ 5000 edges per network.

Then we test stability of each network by measuring its similarity to the network of the previous week. We perform two hypothesis tests using the Jaccard Distance as test statistic: one using the Erdős-Rényi null model and one using the switching randomization null model. We sampled $K = 1000$ networks from each null model, and we estimated a p-value lower than 0.001 that the inferred network of each week is significantly more similar with the network of the previous week rather than the random networks for both null models. In Fig. 5.1 we illustrate the stability of the network for the 12 consecutive weeks for each one of the three inference methods. We use the first week data only to validate the network of the second week.

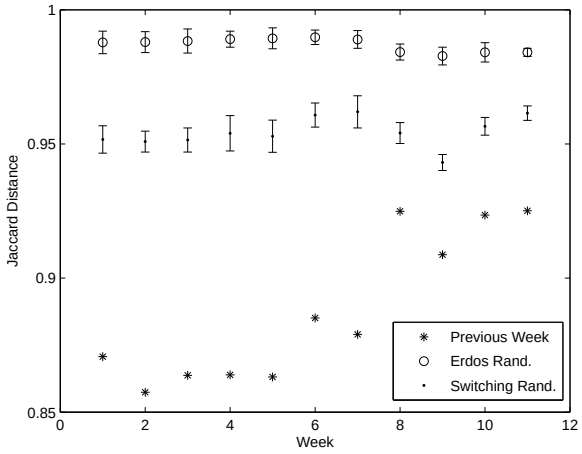
Comparison to Related Networks

In this section we compare the three network inference methods and we validate that their output is related to three reference networks. The reference networks are built using ground truth information of the outlets. In the first reference network, namely 'Location', two outlets are connected if they are based on the same country; In the second, namely 'Language', two outlets are connected if they use the same language; and in the third, namely 'Media-Type', two outlets are connected if they are of the same media type. The three reference networks are formed of several disjoint cliques and we expect that their edges are related to the edges of the inferred network: For

A



B



C

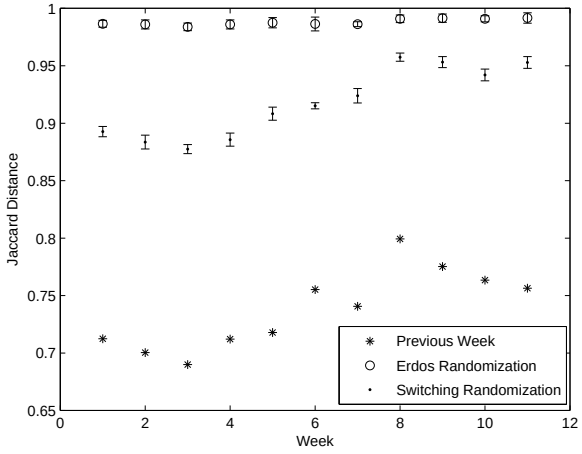
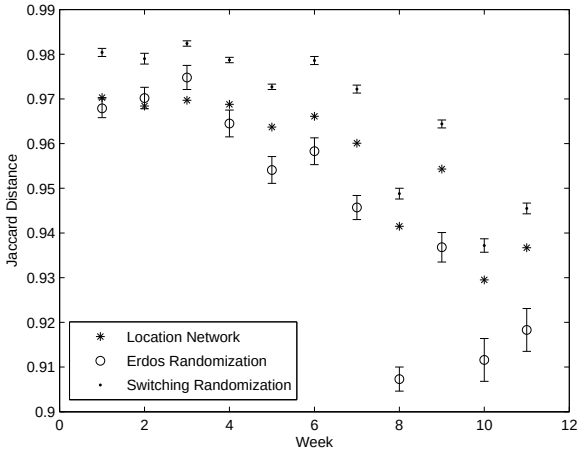
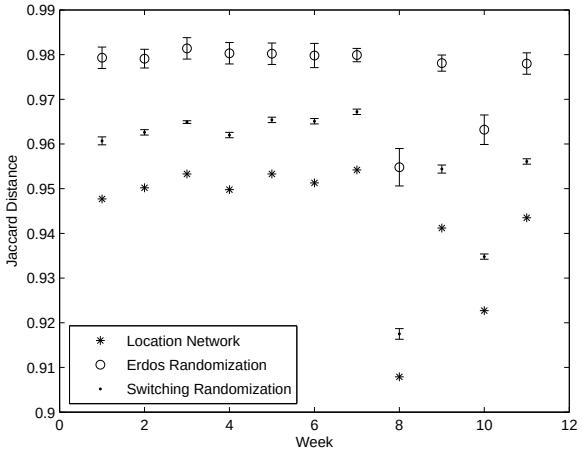


Figure 5.1: Network stability on sequential weeks for the three network reconstruction methods. Errorbars are ±3 standard deviations over the mean value.

A



B



C

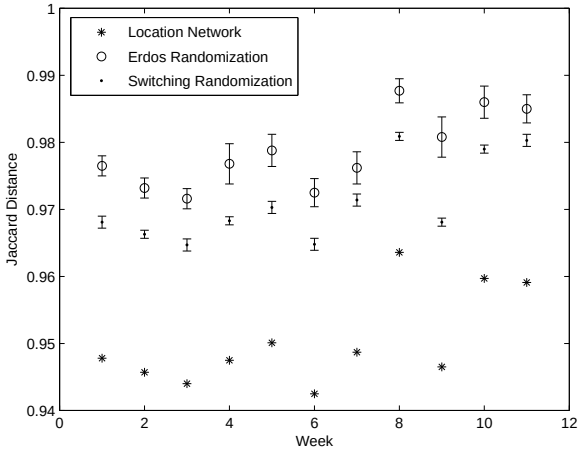
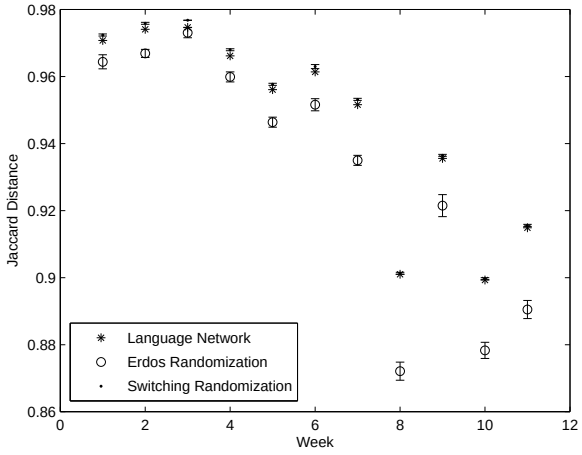
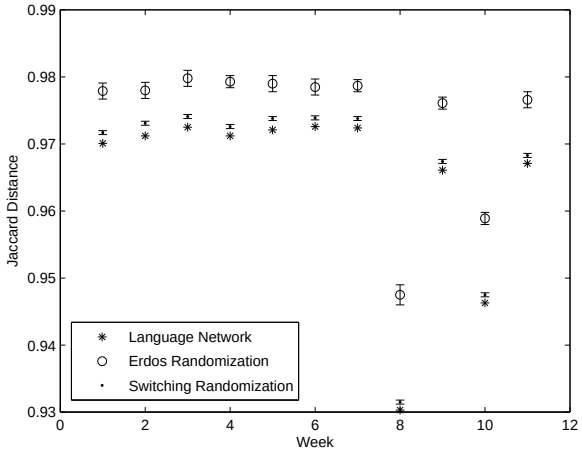


Figure 5.2: Comparison to the 'Location' reference network. Errorbars are set to ±3 standard deviations over the mean value.

A



B



C

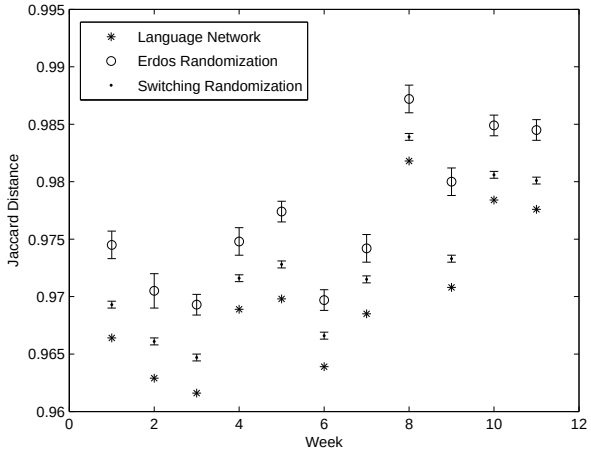
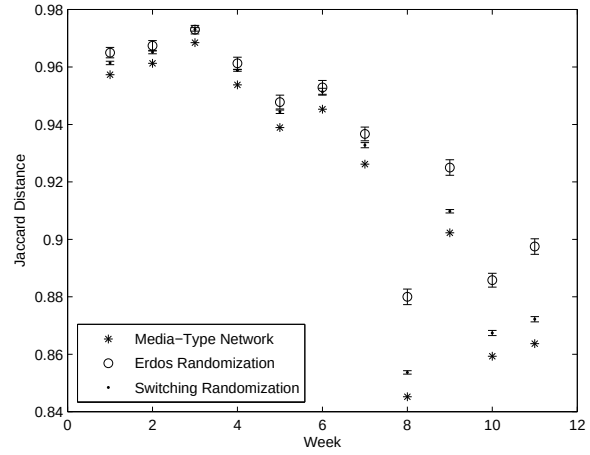
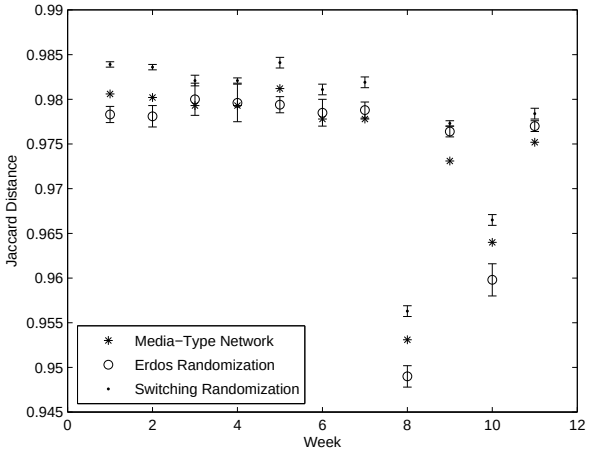


Figure 5.3: Comparison to the 'Language' reference network. Errorbars are set to ±3 standard deviations over the mean value.

A



B



C

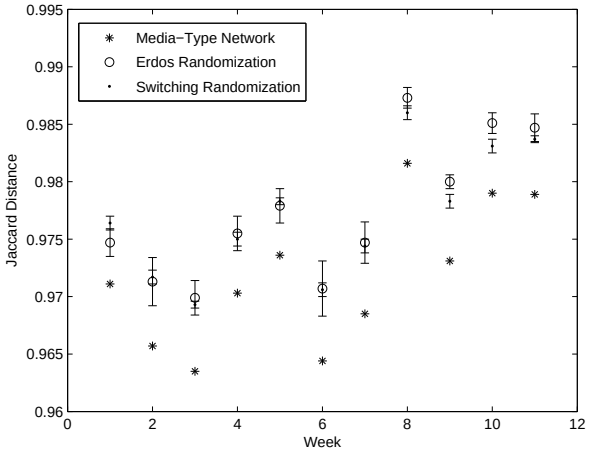


Figure 5.4: Comparison to the 'Media-Type' reference network. Errorbars are set to ± 3 standard deviations over the mean value.

example, a story that is important and publishable in a UK based media probably it is present also in the other UK media, while on the same time it is probably uninteresting to the media of other countries. The influence of the language is also important and it can provide different information compared to the location. For example, we measured that the proportion of articles that mention the word'Pope' in Spanish-language media in the USA is three times larger than the English-language media in the same country. The content of the media is also reflected to their media-type. For example, it is common that stories appear in blogs well before they are published in mainstream media.

We compared the three content-based networks, that are created by the three inference methods, to the three reference networks using the same methodology as before, *i.e.*, the Jaccard distance as test statistic and the two null models. In Fig. 5.2 we compare the three methods to the 'Location' network. Only Methods B and C produce significantly related networks to the reference network for all the 12 weeks ($p < 0.001$). In Fig. 5.3 we present the comparison to the 'Language' network. Only Method C yields significant results for all 12 weeks ($p < 0.001$). Finally in Fig. 5.4 we present the comparison to the 'Media-Type' network where Methods A and C yields significant results ($p < 0.001$). Overall, we can conclude that only Method C produce reliable results over all datasets.

Selecting Inference Method

In the precious sections we compared the three inference methods on their stability in time and on how well their results are related to some ground truth information. We use these results to select the most appropriate method for the news media network. Table 5.1 summarizes the results using the three methods for the 11 weeks' independent datasets and a significance level of 0.001. The best results are achieved using Method C which manages to produce significant results for all the performed tests and all the datasets.

Table 5.1: The number of weeks of a maximum of 11 that each Method presented significant results with $p < 0.001$. ER stands for the Erdő-Rényi Random Graphs and SR for the Switching Random Graphs.

	Previous week		Location		Language		Media-Type	
	ER	SR	ER	SR	ER	SR	ER	SR
Method A	11	11	1	11	0	9	11	11
Method B	11	11	11	11	11	11	2	11
Method C	11	11	11	11	11	11	11	11

Edge Prediction

In this section we will check whether we can predict the existence or absence of edges that are produced by the network inference methods. As described in Sect. 5.1.5 we use a 100-cross-validation scheme to split the inferred network into training and test subnets. We use the training network as a labelled set for the training of a GLM. We will use as features the same three ground truth networks as in Sect. 5.2.2: The Location Network where two outlets are connected if they come from the same country; the Language Network where two outlets are connected if they use the same language; and the Media Type network where two outlets are linked if they are of the same media type. For this experiment we adopted the normal distribution for GLM and the identity link function, *i.e.*, $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. We use the linkage status of the three reference networks as three independent variables $\boldsymbol{X}$; and $\boldsymbol{Y}$ is the corresponding linkage status of nodes in the training subnet. Then we use the trained GLM to predict the linkage status of the nodes in the test subnet. We threshold the corresponding output Y of the GLM: if it is above some threshold we consider the corresponding nodes as connected, otherwise we consider them as unconnected. The selection of the threshold depends on the desired density of the network. In the following experiments we considered several different thresholds that produced networks of different densities. We compare the output of the GLM against the linkage status of nodes as created by the network reconstructions methods. If the reference networks are related to the inferred network we expect that we will be able to predict part of its structure.

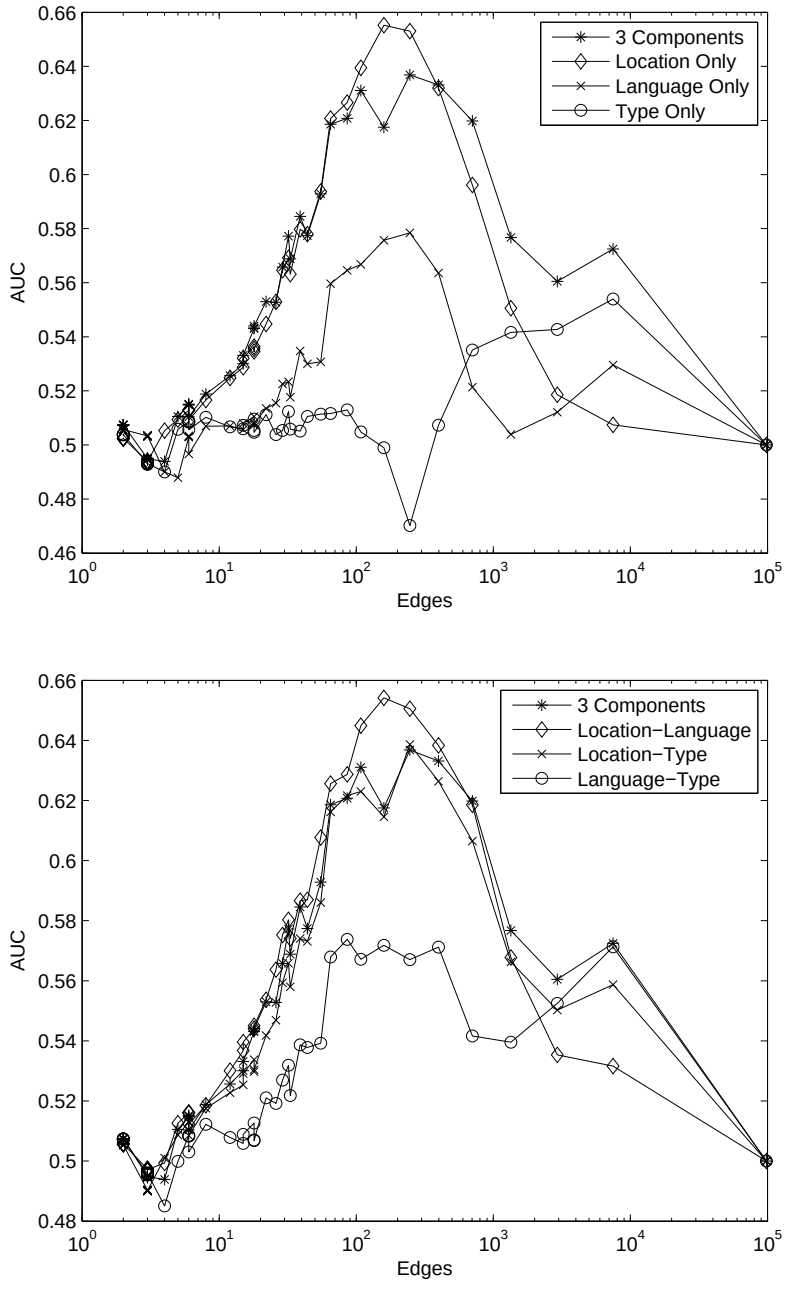


Figure 5.5: AUC accuracy for edge prediction using Method A based on GLM analysis over different network densities,*i.e.*, number of edges.

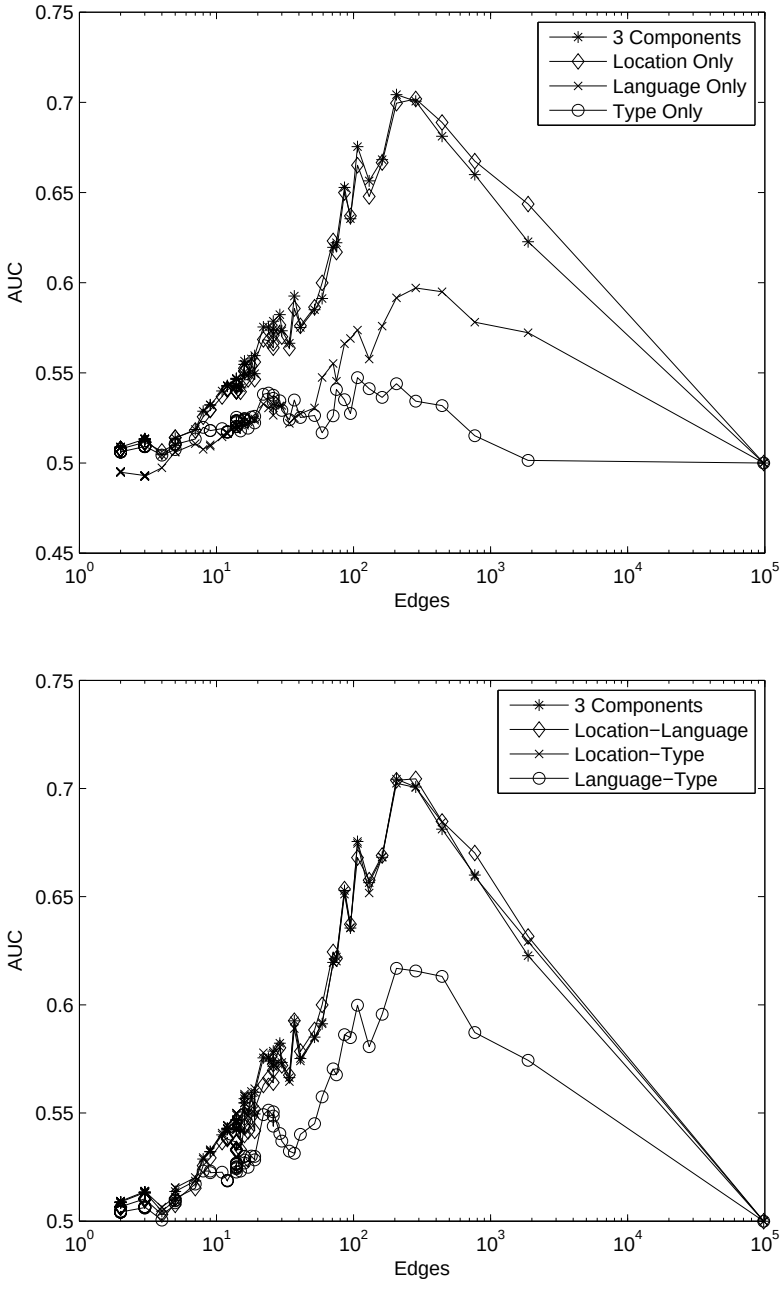


Figure 5.6: AUC accuracy for edge prediction using Method B based on GLM analysis over different network densities,*i.e.*, number of edges.

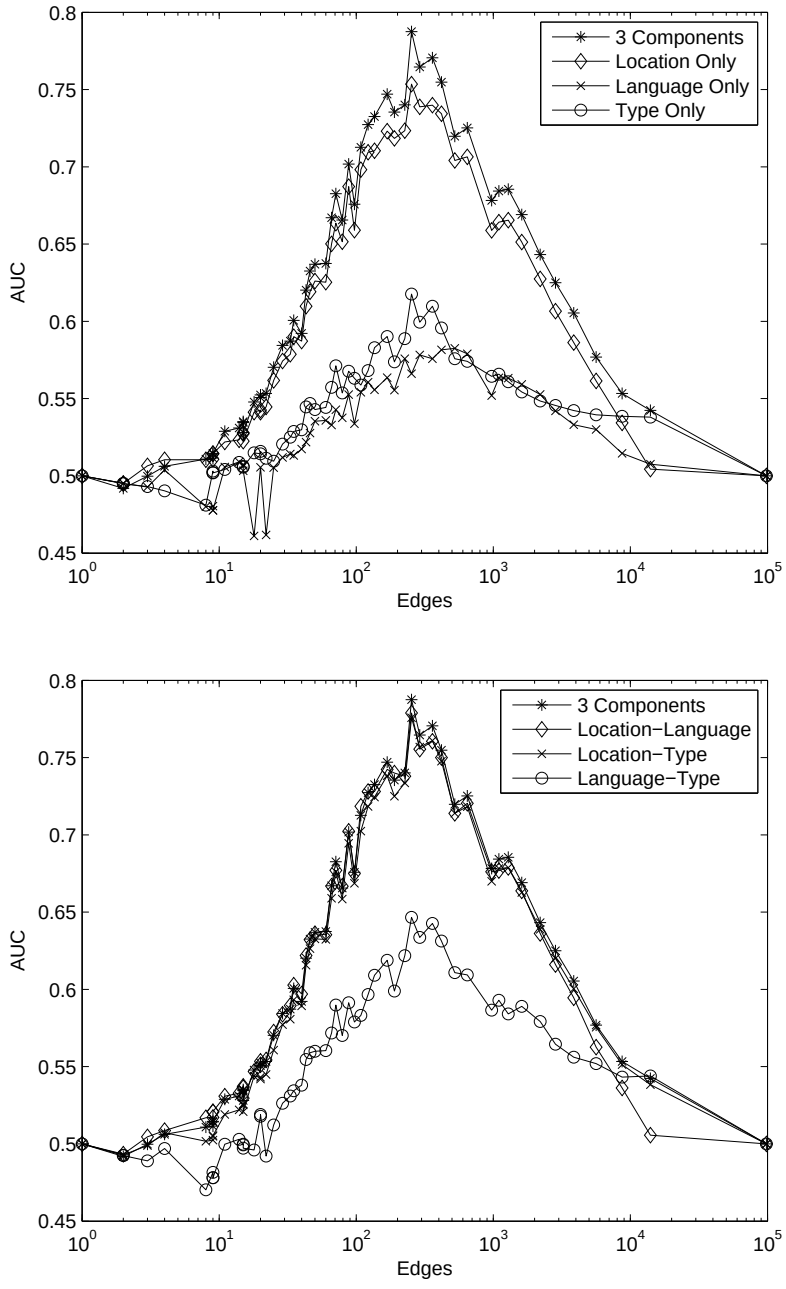


Figure 5.7: AUC accuracy for edge prediction using Method C based on GLM analysis over different network densities,*i.e.*, number of edges.

We measured the AUC for a 100-fold cross-validation scheme for different densities of the outlets network as inferred by the three inference methods using a framework as described in Sect. 5.1.5. Figure 5.5 presents the results for Method A; Fig. 5.6 for Method B; and Fig. 5.7 for Method C. For each method present two sub-figures: In the first we compare the performance of GLM using only one single ground truth network, *e.g.*, only the Location network; and in the second using pairs of ground truth networks. The performance using all three networks is illustrated in both sub-figures two ease comparison. The best prediction over all different network densities reached up to 77.11% AUC using all three ground truth networks and the inference Method C.

5.2.3 Network Visualization

Having decide the most appropriate network reconstruction method, *i.e.*, Method C, we can go a step further and visualize that inferred network from that method. The visualisation is illustrated in Fig. 5.8. To the best of our knowledge this is the first visualisation of the network of media outlets, where outlets are connected based on their interest on publishing the same stories. We used the Cytoscape software to visualise the network and and the more precisely the spring embedding algorithm [163].

A high threshold is set to produce more sparse graph for visualisation reasons. This graph is comprised of 543 nodes, which correspond to outlets, and 4783 edges. Outlets from the same country are coloured the same. We can observe the formation of two dominating clusters: the top one is formed by mostly North American media while the lower one by EU media.

5.3 Summary

In this chapter we introduced a general methodology for validating inferred networks where ground truth is not available. We based our approach on statistical and machine learning methods. We applied the methodology for the validation of the network of news outlets, where outlets are connected

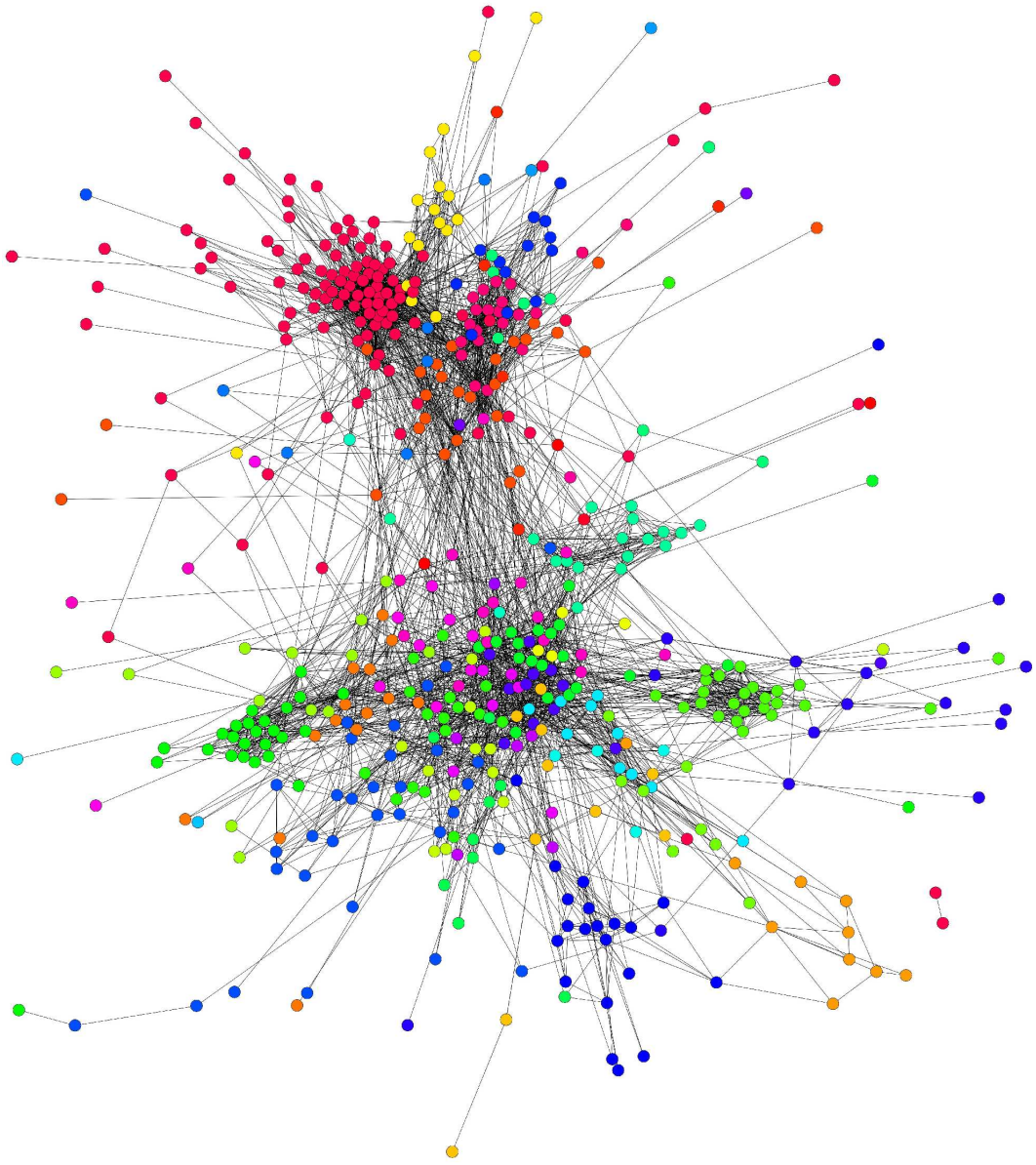


Figure 5.8: A snapshot of the News Media Network. This graph is comprised of 543 nodes, which correspond to outlets, and 4783 edges. We connect two outlets by an edge, if they publish similar stories. We observe the formation of two clusters: the top one is formed by North American media while the lower one by EU media. Outlets from the same country are coloured the same and singleton nodes are omitted.

if they tend to publish the same stories. We visualised this network for the first time.

All the methods presented here can be directly applied to other domains where network inference is used. In [63] we used our approach to predict the topology and illustrate the stability of two networks of EU countries, namely a) the 'citations' network, where two countries are connected if they cite each other in their media content more than average citations and b) the 'co-coverage' network, where two countries are connected if their media cover the same stories more than average co-coverage of stories. The 'co-coverage' network of countries will be discussed in detail in Chapter 6.

Future work includes the analysis of the graph [102, 113], for example the study of its diameter, the degrees of nodes, clustering coefficient, *etc*. Furthermore, future work towards network inference algorithms will focus on directly optimising the stability and the significance of the inference output. In this direction in the next chapter we will use a different network inference algorithm that is based on measuring the statistical independence between outlets.

Chapter 6

Patterns in the EU Mediasphere

In this chapter, we report a large scale content-analysis of cross-linguistic text, by utilising various Artificial Intelligence techniques [64]. We analyse more than a million multilingual articles from the European Union (EU) detecting a clear structure in the choice of stories covered by the various news outlets. We show that the structure is significantly affected by objective geographic, economic and cultural factors among outlets and countries. For example, outlets from countries that share borders are significantly more likely to publish the same stories. Finally we also show that the deviation from the 'average' content of media is significantly correlated with factors such as membership to the eurozone and their wealth measured by Gross Domestic Product (GDP).

An interesting outcome that stems from this research is that media editors, while independently making a series of editorial decisions, *i.e.*, the choices of the news stories they cover, they shape the contents of the EU mediasphere in a way that reflects the geographic, economic and cultural relations of countries. The detection of these subtle signals in a statistically rigorous way is out of the reach of the traditional methods of media scholars. Our research demonstrates the power of the available computational methods for enabling a significant automation of the media content analysis.

6.1 Dataset

Using NOAM, we analysed news articles from the 27 EU countries in 22 languages[1]. For each country we focus on the top-ten news outlets, as ranked by the volume of their web traffic (See Sect. 2.1.1), which offer their content in news feed format suitable for parsing by the NOAM system. For six out of 27 countries we found less than ten outlets with appropriate online presence resulting in a set of 255 outlets. From this set of outlets we managed to successfully parse and analyse a subset of 215 outlets for the period of study. The outlets of the set we monitored are of various types, mainly newspapers and broadcast media. In this set of experiments, in order to make outlets comparable, we used only the news items that were published in the Main feed of each outlet.

Our experiments analyse the contents of those outlets for a period of six months from August 1st, 2009 until January 31st, 2010. In total we analysed 1,370,874 news items. All non-English language news items (about 1.2M) were machine translated into English (See Sect. 2.4.3). All untranslated words were removed before further process in order to minimise the effect of the original language the articles are written to, to the measurement of the similarity between them.

Using the NOAM system, the English and the translated news-items were then processed by typical text mining and Natural Language Processing techniques, *i.e.*, stop word removal, indexing, stemming and TF-IDF representation. Next, articles were clustered into stories by utilising the Best Reciprocal Hit method (See Sect. 2.5) and the cosine similarity. Every outlet was associated with a set of stories and in the next section we will use this information to infer the underline network in a statistically rigorous way.

[1]This study refers to the EU status as on December 2010.

6.2 Experiments

6.2.1 Network reconstruction based on statistical independence

A network reconstruction approach which is based on statistical independence can be realised by utilising the chi-square test statistic. The chi-square test measures statistical independence between two variables [30]. For our application, the first variable is the stories published by outlet A, and the second variable is the stories published by a different outlet B. We run the test for all pairs of outlets. The test concludes whether the two variables are statistically independent, and for our case whether two outlets are independent in their choices of news to cover. If they are not independent we connect them with an edge, otherwise we do not.

The measurement of the chi-square statistic between outlet A and outlet B requires the count of: a) how many stories both outlets published that we denote by S_{11}; b) how many stories A published but B didn't publish that we denote by S_{10}; c) how many stories B published but A didn't publish that we denote by S_{01}; and d) how many stories other outlets published that neither outlet A or B published that we denote by S_{00}. Chi-square requires the computation of the expected counts that A and B would have if they were independent. This computation is given by the following formulas:

$$E_{11} = (S_{11} + S_{01})(S_{11} + S_{10})/N \tag{6.1}$$

$$E_{10} = (S_{10} + S_{11})(S_{10} + S_{00})/N \tag{6.2}$$

$$E_{01} = (S_{01} + S_{11})(S_{01} + S_{00})/N \tag{6.3}$$

$$E_{00} = (S_{00} + S_{01})(S_{00} + S_{10})/N \tag{6.4}$$

where N is the total number of stories.

The chi-square test statistic of outlets A and B is calculated by the quan-

tity:

$$X^2 = (S_{11}-E_{11})^2/E_{11}+(S_{10}-E_{10})^2/E_{10}+(S_{01}-E_{01})^2/E_{01}+(S_{00}-E_{00})^2/E_{00} \tag{6.5}$$

This quantity is associated with the probability that the two outlets are independent. If this probability is above some significance threshold, we consider outlets A and B dependent and we connect them with an edge. The significance threshold is the same for all pairs of outlets. The network inferred by chi-square has a physical meaning: two outlets are connected if they cover the same stories more than expected by chance.

For each pair of the 255 outlets in our dataset we calculated the χ^2-square test statistic and measured their statistical dependence. A threshold was applied to chi-square values to obtain a network of the correlated news outlets. Figure 6.1 illustrates the network of the EU mediasphere outlets. For the chosen threshold value, we obtained 203 non-singleton nodes and 6702 edges among them.

The same approach of chi-square network reconstruction is also applied in Sect. 6.2.3 for the reconstruction a network of countries, where instead of the stories that are published by an outlet we consider the stories that are published in a country.

6.2.2 Communities in the EU Mediasphere

In this section we investigate the underlying structure of the mediasphere network of Fig. 6.1. We based our analysis on the concept of modularity introduces by Newman et al. [142], which measures the quality of the division of a network into modules, clusters of connected nodes. A high modularity reveals a 'good' division where there are many edges between nodes of the same module and few edges between nodes that belong to different connected components. Modularity is calculated based on the formula:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta\left(c_i, c_j\right)$$

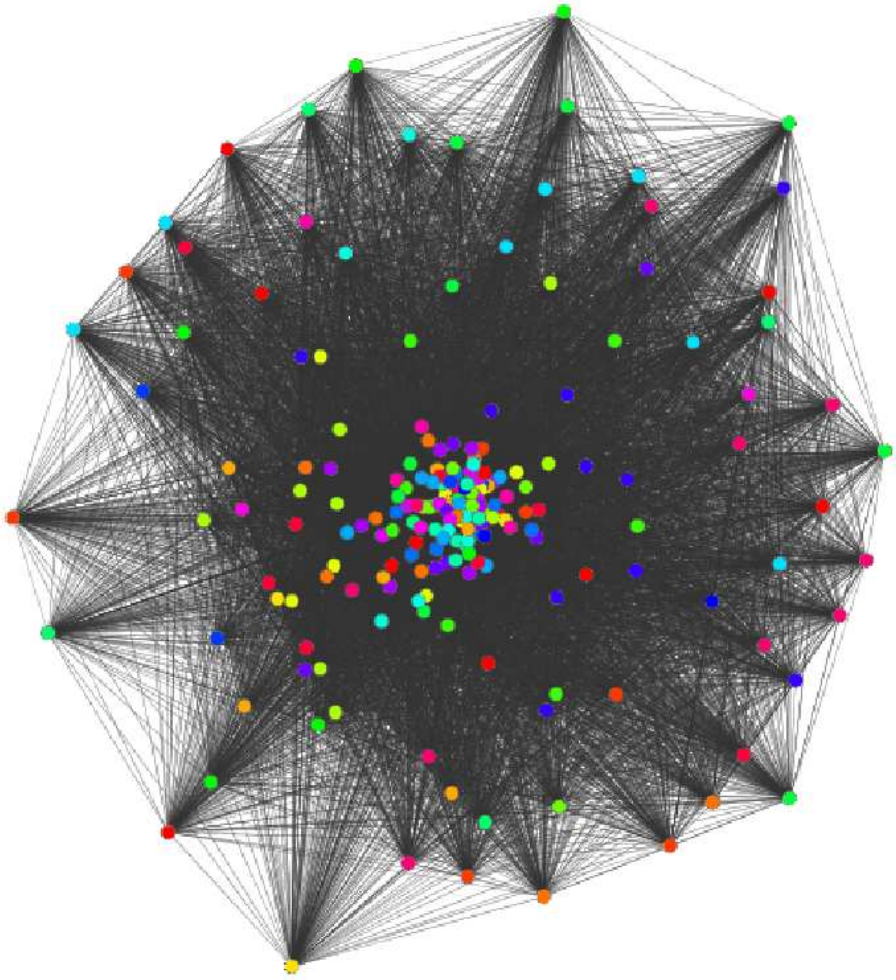


Figure 6.1: A visualisation of the network of EU outlets. This network is comprised by 203 nodes and 6702 edges. Nodes are outlets and edges link outlets that publish the same stories more than expected by chance. Each outlet is coloured by the country of its origin. Disconnected nodes are omitted.

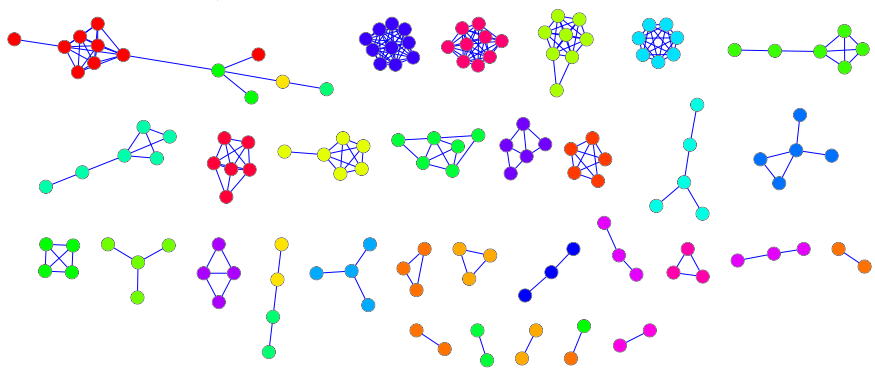


Figure 6.2: The communities of news outlets in the EU mediasphere. This network is a sparser version of the network of Fig. 6.1. Each outlet is coloured by the country of its origin and disconnected outlets are omitted.

where A_{ij} is element (i, j) of the adjacency matrix; m is the total number of edges; k_i the degree of node i; $\delta(c_i, c_j)$ is a delta function; and c_i the connected component to which node i belongs.

We measured the modularity of the network for different significance thresholds T of the χ^2 values. We found that for $T = 21211$ the modularity of the network is maximized and reaches to 0.93. This network is visualised in Fig. 6.2. After the removal of the unconnected nodes, the network is comprised of 147 nodes and 263 edges. The nodes are organised in 31 connected components which roughly correspond to the 27 EU countries. We calculated how probable could this network organisation be by chance, *i.e.*, we measured the probability of two outlets from the same country to end up in the same connected component. This probability is 82.95% and it is significant ($p < 0.001$) as estimated by a randomisation test (by permuting country labels of the outlets and making in total 1000 shuffles).

6.2.3 Analysing Relations among Countries

In the previous section we showed that the network of outlets is decomposed into a set of disconnected components that correspond to the EU countries. This allows us to extent our research and use countries as the 'unit of analysis'

Table 6.1: Factors that affect the choice of stories that are covered in news media.

Factor	Correlation	p-value
Geographical proximity	33.86%	< 0.001
Cultural proximity	32.05%	< 0.001
Trade relations	31.03%	< 0.001

by merging the outlets from the same country. We infer the network of countries by merging the stories that appear in each country in a single set of stories and then by utilising the chi-square approach on those new sets of country-level stories. This time we measure the statistical independence of the media content that appeared in each country compared to the media content in the other countries.

We explored three factors among the many possibilities that may affect the choices of the stories that media in each country choose to cover. These factors are the trade relations between countries, their cultural relations and the existence of common land borders between them. For each of these factors we measured their Spearman correlation to the chi-square scores of the independence of the media content between countries and the results are summarised in Table 6.1:

- To explore the effect of economical relation between countries we used data from United Nations Statistics Division-Commodity Trade Statistics Database in 2008 [2]. We focused on the total of all trade between each pair of countries and more precisely the fraction of the total trade of the country in question that is directed towards each other country. We found a significant correlation of 31.03% ($p < 0.001$) between the trade volume and media content.

- To measure cultural relations between countries we used as data the voting patterns of countries competing in the Eurovision song contest from 1957 to 2003[3]. More precisely we used the fraction of the total

[2]COMTRADE: http://comtrade.un.org/db
[3]Eurovision Song Contest: http://www.eurovision.tv

points awarded by the country in question to each other country over the whole period of time. Countries that participated in the contest and they are not in the current EU countries list were removed prior to normalisation. We found a significant correlation of 32.05% ($p < 0.001$) between the voting patterns and the media content.

- To measure geographical proximity we used the proportion of length of the common land borders between countries. We used this approach instead of binary relations to take into account differences of countries such as for example Portugal which shares borders only with Spain and Germany that shares land borders with many more countries. We found a significant correlation of 33.86% ($p < 0.001$).

If we threshold the chi-square scores between countries, we can get a network of 'relations' between them. We refer to this network as the 'co-coverage' network and we present one illustration of it in Fig. 6.3. That network was built as sparse as possible with the constraint that the network must remain connected. The inspection of the network reveals a series of well understood connections. For example, there are connections between Greece and Cyprus, Czech Republic and Slovakia, Latvia and Estonia, United Kingdom and Ireland, Belgium and France, *etc*. Other connections which are less understood and the reasons that these connections exist could potentially be the basis of further research from social scientists.

6.2.4 Ranking of Countries

In previous section we built a matrix of chi-square scores between each pair of countries. If we treat these scores as similarities we can apply non-metric Multidimensional Scaling (MDS) and embed them in a 26 dimensional space – since we have $N = 27$ points projected in $N - 1$ space. In that space it is easy to compute the centre of mass of all countries and measure the Euclidean distance of the centre from each country separately. The centre of mass represents the average media content in the EU media, and the distance represents the deviation of each country from the average media content. We

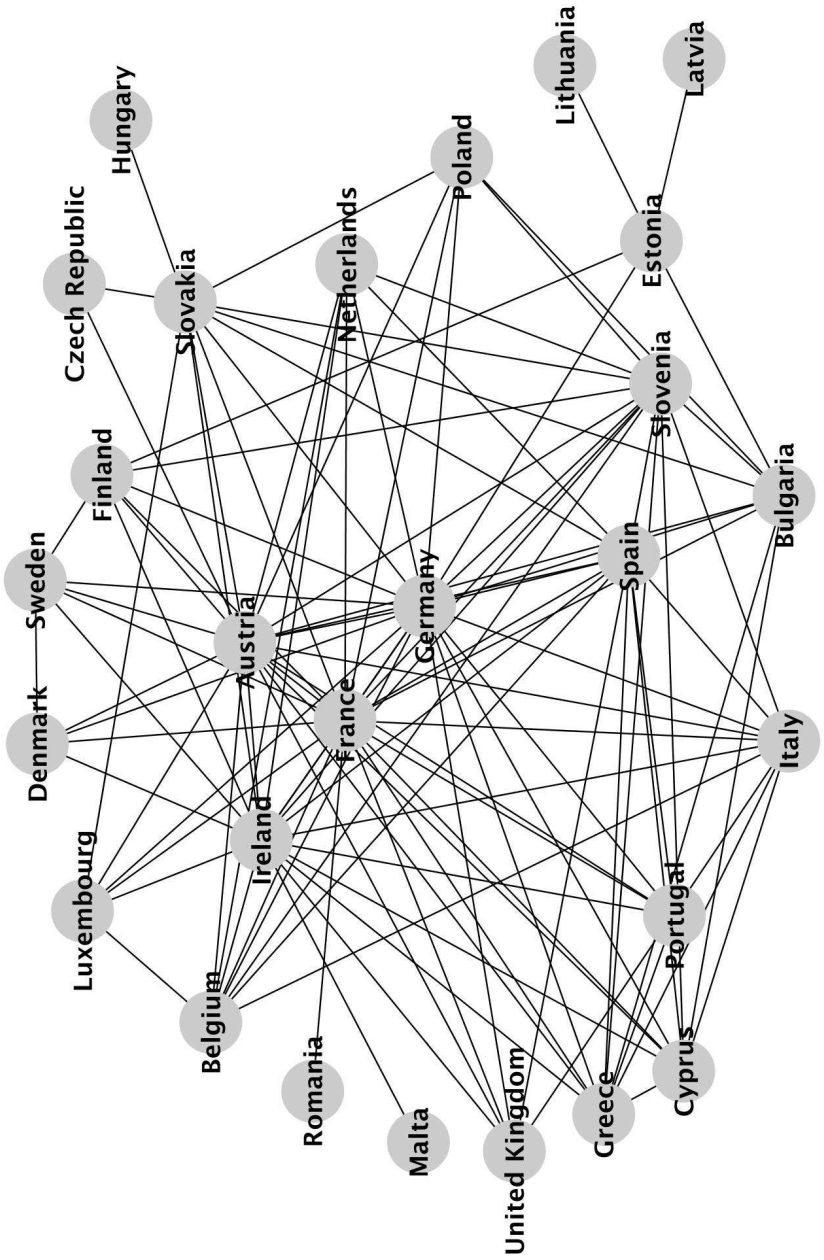


Figure 6.3: The co-coverage network of EU countries where we connect two countries if their media have a significant preference of covering the same stories. The network is comprised of 27 nodes, one node for each EU country, and 112 edges between them. The sparseness of the network was chosen as high as possible with the constraint that the network must remain connected.

Table 6.2: Ranking of countries based on how close their media content is to EU average media content.

Rank	Country	Distance	In Eurozone	Accession Year
1	France	0.6109	Y	1957
2	Austria	0.6161	Y	1995
3	Germany	0.6270	Y	1957
4	Greece	0.6304	Y	1981
5	Ireland	0.6347	Y	1973
6	Cyprus	0.6356	Y	2004
7	Slovenia	0.6505	Y	2004
8	Spain	0.6521	Y	1986
9	Slovakia	0.6558	Y	2004
10	Italy	0.6579	Y	1957
11	Belgium	0.6580	Y	1957
12	Luxembourg	0.6611	Y	1957
13	Bulgaria	0.6639	N	2007
14	Netherlands	0.6668	Y	1957
15	United Kingdom	0.6717	N	1973
16	Finland	0.6755	Y	1995
17	Sweden	0.6763	N	1995
18	Poland	0.6769	N	2004
19	Estonia	0.6774	N	2004
20	Denmark	0.6777	N	1973
21	Portugal	0.6778	Y	1986
22	Malta	0.6846	Y	2004
23	Czech Republic	0.6867	N	2004
24	Romania	0.6880	N	2007
25	Latvia	0.6915	N	2004
26	Hungary	0.7044	N	2004
27	Lithuania	0.7075	N	2004

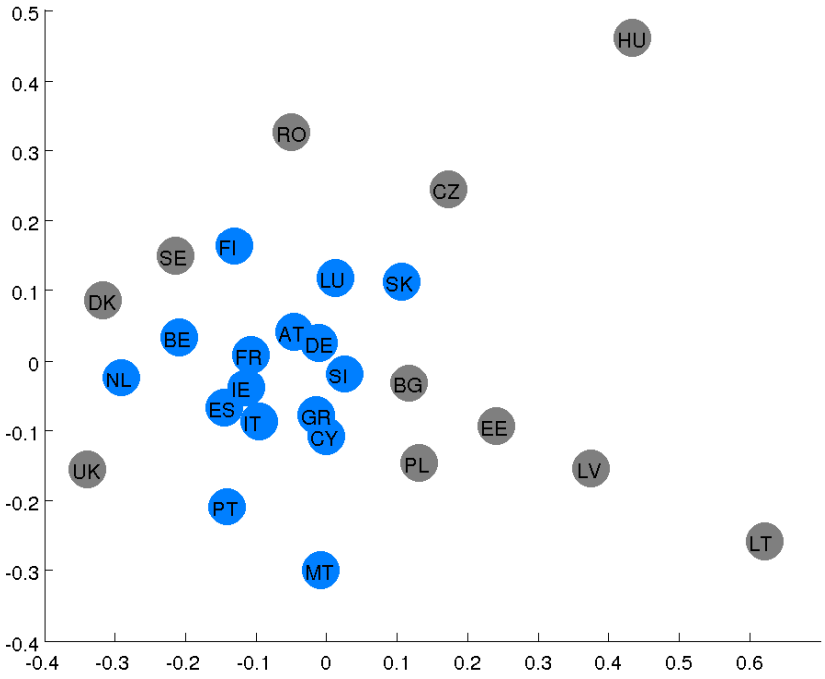


Figure 6.4: The 'co-coverage' map of the EU. The relative positions of the EU countries represent the content similarity of their media content. The countries that use Euro are coloured in blue and the rest are coloured in gray.

ranked the countries based on their distance to the average content and the result is presented in Table 6.2. We found some more patterns that emerge from this ranking. There is a significant correlation 48.94% ($p = 0.0096$) of the ranking to their year of accession of in EU. Also, in the top of the list are mostly found the Eurozone countries. One interesting point is that while UK and Ireland share the same language and have common borders, the news coverage in Ireland seems to be closer to the average EU media content compared the more Eurosceptic UK.

6.2.5 A Media Content based Map of the EU

Following a similar approach, as in the previous section, we project the EU countries into a 2D space using MDS as illustrated in Fig. 6.4. In this projec-

Table 6.3: Correlations of countries deviation from average EU media content and their demographic data.

Factor	Correlation (%)	p-values
Member of Eurozone	70.65	<0.001
Accession Year	-48.94	0.009
GDP 2008	44.75	0.020
Population	23.05	0.247
Area	15.63	0.435
Population Density	7.45	0.712

tion we can observe the relative positions of the EU countries. Countries that share interest in the same stories are located closer together, while countries with less common interests in their media content are located further apart. We can observe that Eurozone countries are closer to each other, and also closer to the origin of the space which represents the 'average' content of the EU media.

Similarly we explored the correlations of the deviation of countries from average EU content and some of their demographic data. These correlations are illustrated in Table 6.3. We found significant correlation to the GDP of each country and no significant correlation to factors such as the size of country, its population or population density.

6.3 Summary

In this chapter, we presented what we believe is the first large scale content-analysis of multilingual news in the social sciences, by deploying various methods including recent advances in statistical machine translation and text analysis. We analysed the content of news media outlets based in the European Union, *i.e.*, the European Mediasphere, in order to detect some emerging patterns. Our approaches not only allow us to examine large scale samples of news articles, but they are also data-driven and allow the data itself to reveal the underline patterns.

Our results are based on the automated content analysis of a corpus of

1.3M articles from all EU countries. Although our methods are less accurate compared to the 'coding' that human coders would provide, they can provide statistically significant results due to the large scale of the data used. Our key findings include:

- The inference of the network of EU leading outlets based on their choices of the news they cover.
- The underline structure of the network are connected components of media from the same country with probability of two outlets from the same country to end up in the same connected component equal to 82.95%.
- The structure of the network is significantly affected by objective geographic ($c = 33.86\%$), economic ($c = 31.03\%$) and cultural ($c = 32.05\%$) factors among countries.
- The network of countries based on the coverage of the same stories in their media.
- The ranking of countries based on how their media deviate from the average EU media content.
- The correlations of the deviation of countries from average EU media content and some of their demographic data: Member of Eurozone (70.65%), Accession Year (-48.94%) and GDP (44.75%).

Finally it is worth mentioning that our approach shows the feasibility of a global scale analysis of the mediasphere by automated means. Also the data driven approach is a significant breakthrough in the news media content analysis.

Chapter 7

Conclusions

This research presented a series of findings that resulted from the automated analysis of news: we compared news outlets based on the topics they select to cover; we measured the linguistic subjectivity and the readability among different topics and among different outlets; we showed the potential of predicting which articles can become popular before they are published; we inferred the network of "outlets that discuss the same stories" and we showed that is it is stable in time and with a predictable structure; we measured how factors such as geographic proximity, economic and cultural relations between countries affect the selection of stories that their media choose to cover.

A great deal of this study's interest focused on the development, benchmark and demonstration of automated approaches in order to answer questions similar to those asked by media scholars. However, this study presented results, such as the inference of the news media network, that go beyond the reach of approaches that are currently used by social scientists. The findings presented in this study suggest not only that research can now be conducted across multiple languages and countries, but also that quantitative methods and techniques can be used to study media system on a global scale. Indeed, experiments can be conducted that include millions of news articles published for long time periods. Having presented the findings of this study, the present chapter discusses some general implications of our research while

it also presents some potential avenues for future research.

As previously discussed, our methods open access to large scale analysis, and by exploiting the law of large numbers, they can reach statistically significant results. Furthermore, we presented experiments that could not, for practical reasons, be conducted without the deployment of automated methods, *e.g.*, the analysis of multilingual corpora. Nevertheless it is highly acknowledged that the interpretation of the results remains a field of research that can be only conducted by human analysts, since the statistical methods are currently limited to 'shallow semantics'. There is no argument that shallow semantics may offer more information than simple string matching techniques, however, they still miss the real meaning of articles. On the contrary human coders can be more accurate than automated approaches and offer better interpretations of the results, they are restricted to corpora of limited size. In parallel, it is worth mentioning that human coders suffer from subjectivity when they code texts and may introduce their own bias to the research, something not applicable to automated approaches.

This study also presented a potential impact on the research of other domains of social sciences because of the automation of media studies. For example in Chap. 6 we presented how the geographical, cultural and trade relations of countries are reflected in the stories that their media choose to cover. The study of relations of countries goes beyond the scope of research of media scholars and it is a field of research of political sciences [103]. A different domain that could be affected is finance, since there are works that reveal that the monitoring of the sentiment of media content can predict financial markets [196, 127].

7.1 Future Work & Open Questions

We anticipate that the present study may yield potential fruitful avenues for future research. These potential opportunities may be divided into two main categories. The first, incorporates the theories of media studies that could, at least partially, be confirmed by utilising large scale news analysis methods while the second concerns technical questions that emerge from our research.

7.1.1 Questions on Media System

We argue that solving real problems that directly emerge from the Social Sciences should be a the goal of Computational Social Science – rather than adapting the problem to what can actually be solved by computational methods. Towards this direction, we focus on two theories which have been developed by media scholars. The first is in regards to the news values which make a story newsworthy and the second concerns the 'filters' theory of Herman and Chomsky [89].

News values

Several works on media studies aim to identify the news' values that make a story newsworthy for a given audience [71]. These theories identify some criteria that will identify, for a given outlet and for a given audience, whether a report is worth to be published. For example, Hetherington pointed out several values including significance, drama and surprise [90]. Galtung and Ruge identified 12 news values [74]. A recent study by Harcup and O'Neill [83] was based on the aforementioned study and suggests the following: reference to some elite individual, organisation or nation; entertainment (*e.g.*, drama, sex); good news (*e.g.*, triumphs, rescues); bad news (*e.g.*, disasters, tragedy); relevance (*e.g.*, proximity of audience to the event, cultural proximity) magnitude; surprise; follow-up stories; and the general political agenda of the outlet. We believe that most of these criteria could be modelled and tracked. Thus, a system could be built to identify the newsworthy stories. This system could be of value for editors as it can provide a quick screening of the large volume of press releases that they have to explore in order to find the 'best' stories; and even more importantly it could be of value for Public Relations offices in order to have an estimation of what are the chances of their press releases being picked up by media.

Filters Theory

A 'healthy' media system is expected to provide access to a diversity of stories and opinions. However, it is unclear if this is the case nowadays despite the

vast number of news outlets that can be reached relatively easily and freely. In their seminal work Herman and Chomsky [89] propose five 'filters' that are expected to influence news that appear in the media: the ownership of the media; the use of the same common sources from journalists; advertising; 'flak' defined as an organised attempt to control media content; and fear of a common enemy. We argue that at least the first three filters could nowadays be experimentally checked in a quantitative manner. For the first filter, we need to record the ownership of each outlet and then check for relation of this factor to the actual media content. It is expected that some stories will have a different presentation angle (sentiment) compared to some average or they can even be completely missing. For the advertising filter we need to record the advertisements that appear in each outlet and follow the links to actual companies that pay for the advertisements. Then a similar checking can be performed as in the case of ownership. The third filter requires the tracking of media content to their primary sources such as press releases or news agencies. It is known that most online media content about international news can be tracked to the same sources [146, 118]. The text reuse check can be implemented by utilising methods such as a suffix tree [166].

7.1.2 Technical Questions

Our research has lead to some technical questions that the current literature fails to answer. We discuss two of those questions: The modelling of news article in a kernel-based manner and the study of the effect of Machine Translation to classification or clustering.

News-items as Structured Objects

News items are structured objects as illustrated in Table 2.2. A news article is represented by a set of information including title, description, content, date, outlet source, related topic tags, set of entities such persons, organisations and places, and so on.

The study of this kind of structured objects has a long history in AI community. Several fields of machine learning and data mining offer methods

and tools for the direct analysis of structured data, including Inductive Logic Programming [43] and Multi-Relational Data Mining [47] to name a few.

Kernel based methods, aim to model similarity between news items in a more sophisticated manner than the simple cosine similarity between the concatenation of their titles, description and content: For each one of these elements a different kernel function can be defined that measures their similarity. For example similarity between titles could be measured by a cosine kernel function and between dates by Gaussian kernel. Linear combinations of these kernels, in terms of kernel engineering, can lead to the generation of sophisticated kernels that measure better the similarity of articles in a uniform way. The combination of kernels in order to model similarities of structured objects is an active research field [177, 168, 11, 109, 110].

The Effect of SMT on Supervised and Unsupervised Learning

Another issue that emerges from our research is the effect of Statistical Machine Translation to supervised and unsupervised learning. Both problems ask whether and how the geometry of the vector space is affected by SMT. Is it the same the result of a clustering of articles in their native language and in their machine translated version? Does a tagger that is trained in English work as expected to a machine translated input? We conducted some preliminary experiments that show that there is a correlation of the machine translation quality in relation to the performance of taggers. There are also indications from the information retrieval community that SMT does not significantly affect the query process [158]. Nevertheless, this is still an open problem in the literature.

7.2 Data-Driven Research

A final comment for discussion concerns the current approaches of conducting research in social sciences, as well as in media studies, which are fully hypothesis-driven. Currently, the social scientists formalise a set of questions and with the use of manual 'coding' approaches they collect and analyse the

data in order to conclude regarding the validity of the predefined hypothesis. Thus, they can answer only those specific research questions on which they focus. An important step in the research is to pose the correct questions that will lead to some meaningful results: they have to observe the collected data and detect the underlying potential patterns and trends. These patterns help them to formalise the hypotheses they will further investigate.

However, given a rich collection of data, this hypothesis-driven research is highly restricted since only trends or patterns for which there is a suspicion of presence can be discovered. In a massive dataset, there may be present significant patterns which could be missed by the researcher. Furthermore, there is a high cost to potential changes to the research questions after the hypotheses have been posed and the coding process has started, since even a minor change in a current hypotheses or the addition of a one would result in a re-coding of the whole dataset.

Nowadays, the plethora of data and the automation of research can lead towards a 'data-driven' research approach. This makes possible for patterns to emerge from the data, rather than detected simply in response to a coding frame. This has already happen to various other disciplines, including physics and biology [50, 151, 9, 78, 94]. Social sciences can also be impacted, particularly those disciplines concerned with the analysis of text, due to the recent availability of millions of books and news articles in a digital format [136, 73, 38]. In this direction, in our research we applied 'data-driven' approaches in Chap. 5 where we inferred the news media network; and in Chap. 6 where we studied the relations between countries as they are reflected in their media content. For those results no formal hypothesis was set *a priori*. In the first case we found the significant relations between news outlets and in the second the significant relations between countries.

Chapter 8

Appendices

8.1 Supplementary Tables

Table 8.1: Number of Outlets and Feeds in NOAM by Location

Location	Outlets	Feeds	Location	Outlets	Feeds
Afghanistan	2	2	Peru	9	9
Africa	4	17	Philippines	2	9
Albania	1	1	Poland	26	26
Algeria	8	8	Portugal	25	27
Argentina	6	7	Qatar	1	2
Armenia	4	4	RepMacedonia	1	1
Asia	5	5	Romania	17	17
Australia	17	81	Russia	7	7
Austria	17	17	Rwanda	1	1
Bahamas	1	1	Saudi Arabia	1	1
Bahrain	4	8	Senegal	2	2
Bangladesh	11	16	Serbia	2	2
Belgium	12	15	Sierra Leone	1	1
Belize	2	2	Singapore	1	6
Bhutan	3	3	Slovakia	9	9
Bolivia	4	4	Slovenia	8	12
Bosnia-Herzegovina	1	1	Somalia	4	4
Botswana	5	5	South Africa	12	21
Brazil	8	15	South Korea	1	1

Location	Outlets	Feeds	Location	Outlets	Feeds
Bulgaria	17	17	Spain	19	22
Cambodia	2	2	Sri Lanka	8	8
Cameroon	3	3	Sudan	1	1
Canada	29	63	Swaziland	1	1
Chile	6	6	Sweden	46	50
China	7	23	Switzerland	1	4
Colombia	10	11	Taiwan	1	2
Congo	1	2	Thailand	2	2
Costa Rica	6	6	Togo	2	2
Croatia	2	2	Trinidad and Tobago	1	1
Cuba	10	10	Tunisia	1	1
Cyprus	7	10	Turkey	7	17
Czech Republic	15	18	UAE	3	18
Denmark	23	23	Uganda	7	7
Ecuador	10	10	Ukraine	3	3
Egypt	2	2	United Kingdom	96	336
El Salvador	3	3	United States	591	1387
Estonia	16	21	Uruguay	3	3
Ethiopia	2	4	USA Alabama	0	1
Finland	34	34	USA Alaska	0	1
France	27	58	USA Arizona	0	8
Gambia	5	6	USA California	0	43
Georgia	2	2	USA Colorado	0	8
Germany	32	66	USA Connecticut	0	3
Ghana	4	6	USA Delaware	0	5
Greece	31	42	USA Florida	0	7
Greenland	1	1	USA Georgia	0	4
Guatemala	3	3	USA Hawaii	0	19
Guyana	5	5	USA Idaho	0	2
Haiti	2	2	USA Illinois	0	15
Honduras	5	5	USA Indiana	0	5
Hong Kong	1	2	USA Iowa	0	20
Hungary	19	22	USA Kansas	0	9
Iceland	1	1	USA Kentucky	0	1
India	14	33	USA Louisiana	0	2
Iran	4	14	USA Maine	0	1
Iraq	1	1	USA Maryland	0	6
Ireland	13	26	USA Massachusetts	0	21
Israel	9	15	USA Michigan	0	18
Italy	32	69	USA Minnesota	0	10

Location	Outlets	Feeds	Location	Outlets	Feeds
Jamaica	5	5	USA Mississippi	0	2
Japan	5	14	USA Missouri	0	1
Jordan	1	1	USA Montana	0	2
Kazakhstan	2	2	USA Nebraska	0	5
Kenya	5	6	USA Nevada	0	2
Kuwait	1	1	USA New Hampshire	0	1
LatinAmerica	2	7	USA New Jersey	0	1
Latvia	8	9	USA New Mexico	0	2
Liberia	1	1	USA New York	0	76
Liechtenstein	2	2	USA North Carolina	0	4
Lithuania	12	13	USA North Dakota	0	1
Luxembourg	8	8	USA Ohio	0	13
Madagascar	1	1	USA Oklahoma	0	3
Malawi	1	1	USA Oregon	0	1
Malaysia	2	4	USA Pennsylvania	0	20
Maldives	4	4	USA Rhode Island	0	1
Malta	6	7	USA South Carolina	0	6
Mauritius	5	5	USA South Dakota	0	1
Mexico	13	19	USA Tennessee	0	9
Middle East	3	7	USA Texas	0	21
Monaco	1	1	USA Utah	0	4
Morocco	4	4	USA Vermont	0	1
Namibia	3	3	USA Virginia	0	13
Nepal	3	3	USA Washington	0	24
Netherlands	37	37	USA West Virginia	0	1
New Zealand	3	17	USA Wisconsin	0	12
Nicaragua	5	5	USA Wyoming	0	2
Nigeria	16	16	Uzbekistan	1	1
Norway	1	1	Vatican	2	3
Pakistan	5	10	Venezuela	4	7
Panama	6	6	Yemen	2	2
Paraguay	3	3	Zimbabwe	16	16

Table 8.2: List of outlets in NOAM

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
13wham.com	3	leidschdagblad.nl	1
168ora.hu	1	lejournal.euskalherria.com	1
20minutes.fr	1	lejournaldetanger.com	1
20minutos.es	1	lejsl.com	1
24zimbabwe.com	1	lejsl.fr	4
2snaps.tv	1	lemonde.fr	10
2theadvocate.com	2	lenouvelliste.com	1
5septiembre.cu	1	leparisien.fr	5

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
a1nigerianews.com	1	lequotidien.editpress.lu	1
aachener-zeitung.de	6	lequotidienalgerie.com	1
aamulehti.fi	1	lesechos.fr	4
aawsat.com	1	lesoir.be	1
abc.net.au	4	lesoleil.sn	1
abc4.com	1	lessentiel.lu	1
abcnews.go.com	9	leta.lv	1
abdn.ac.uk	2	letelegramme.com	3
abendblatt.de	6	levante-emv.com	1
aberdeennews.com	1	lexpress.fr	1
abertay.ac.uk	1	lexpress.mu	1
abeshabunnabet.com	1	lexpressiondz.com	1
abqjournal.com	1	lhv.ee	1
accessatlanta.com	1	libdems.org.uk	1
accra-mail.com	1	liberation.fr	3
aclu.org	1	liberianobserver.com	1
acm.org	1	libero-news.it/	1
acorianooriental.pt	1	libertaddigital.com	1
actualitatea-romaneasca.ro	1	libertatea.ro	1
actualno.com	1	lidovky.cz	1
ad.nl	1	lifeinitaly.com	1
addisonindependent.com	1	limburger.nl	1
AdelaideNow	8	limerickleader.ie	1
adelante.cu	1	link.brightcove.com	1
adevarul.ro	1	littlegreenfootballs.com	1
admin.cam.ac.uk	1	littlehamptongazette.co.uk	1
adnkronos.com	4	ln-online.de	5
advertiser.ie	1	lodinews.com	1
aeiou.expresso.pt	1	lokalavisen.dk	1
aerztezeitung.de	3	londoncranes.com	1
affaritaliani.it	1	lostiempos.com	1
afp.com	1	lr-online.de	1
africasciencenews.org	1	lrt.lt	1
afterellen.com	1	lrytas.lt	1
aftermathnews.wordpress.com	1	ls24.fi	1
aftonbladet.se	1	lsus.edu	1
againsthillary.com	1	luc.edu	1
agi.it	5	luxpost.editpress.lu	1
ahora.cu	1	lyonne-republicaine.fr	3
aina.org	1	maaleht.ee	1
airamericaradio.com	1	maaseuduntulevaisuus.fi	1
aja.com.pe	1	madagascar-tribune.com	1
ajc.com	3	madata.gr	1
akenyanews.com	1	madison.com	2
aktualne.centrum.cz	1	maerkischeallgemeine.de	3
alalam.ir	1	magyarhirek.hu	1
albanianeconomy.com	1	maiahoje.pt	1
aldia.cr	1	makemeheal.com	1
aldiatx.com	4	makfax.com.mk	1
alfa.lt	1	malaysiasun.com	1
algarveobserver.com	1	maldivesreports.com	1
aljazeera.net	2	malmo.se	1
allafrica.com	12	maltamediaonline.com	1
allehanda.se	2	manchestereveningnews.co.uk	1
allheadlinenews.com	1	manchesteronline.co.uk	1
almasryonline.com	1	marca.com	1
ambito.com	1	marietta.edu	1
ameinfo.com	15	marketwatch.com	1
americandaily.com	1	marylhurst.edu	1
americanthinker.com	1	mauinews.com	1
amnesty.org	1	mauritius-news.co.uk	1
amtsavisen.dk	1	mbs.ac.uk	1
an-online.de	7	mcall.com	1
andrewsullivan.theatlantic.com	1	mdn.mainichi.jp	5
ansa.it	5	me-ontarget.com	1
ant1.com.cy	1	medheadlines.com	1
antioch-college.edu	1	mediafax.ro	1
antiwar.com	1	mediamonitors.net	1
aok.dk	1	mediatimesreview.com	1
ap.hellomagazine.com	1	medicalnewstoday.com	2
ap.org	17	mediterraneonews.it	1
ap3.ee	1	MeetBarackObama.com	1
apcom.net	1	meforum.org	1
apfanews.com	1	megatv.com	1
apogevmatini.gr	1	meiemaa.ee	1
aps.org	2	memphisdailynews.com	1
arabtimesonline.com	1	menara.ma	1
arbeiterkammer.at	1	merkur-online.de	1
arbetarbladet.se	1	messenger.com.ge	1
arbor.edu	1	messiah.edu	1
archive.gulfnews.com	2	metro.co.uk	3
armeniadiaspora.com	1	metro.cz	1
armenianow.com	1	metro.se	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
army.mil	2	metrofrance.fr	7
asianage.com	5	mfdnes.cz	1
asianews.it	1	mg.co.za	1
asiantribune.com	1	miadhu.com.mv	1
asiasentinel.com	1	miamiherald.typepad.com	1
associatedcontent.com	7	michaelmoore.com	1
aston.ac.uk	1	michellemalkin.com	1
astronomy.com	1	microsoft.com	2
asu.edu	1	midilibre.com	4
atarde.com.br	1	minivannews.com	1
athensnews.com	5	minnhalsafi.blogspot.com	1
athina984.gr	1	mirror.co.uk	3
atimes.com	1	missouristate.edu	1
auniao.com	2	missouristate.edu	1
austin360.com	1	mit.edu	5
austriacreditscoreonline.co.uk	1	mittelbayerische.de	1
austrianinformation.org	1	mladina.si	1
austriantimes.at	1	mlive.com	4
autosport.com	3	mlive.com/cafe/	1
az.com.na	1	mmegi.bw	1
azcapitoltimes.com	1	mno.hu	1
azcentral.com	6	mnsu.edu	1
azstarnet.com	2	modernghana.com	3
b92.net	1	mollygood.com	1
backtorockville.typepad.com	1	monaco.mc	1
baden-online.de	1	monde-diplomatique.fr	1
ballard.co.uk	3	mondediplo.com	1
balticbusinessnews.com	1	monitor.co.ug	1
baltictimes.com	1	montana.edu	1
baltimoresun.com	11	montclair.edu	1
baltimoresun.com	1	morgenpost.de	7
baltische-rundschau.eu	1	morgenweb.de	6
banateanul.ro	1	morningstar.com	1
bangkokpost.com	1	msnbc.msn.com	16
bankofengland.co.uk	1	msu.edu	1
barackobama.com	5	mtairynews.com	1
barneveldsekrant.nl	1	mtu.edu	1
barometern.se	1	mtv.com	1
barricada.com.ni	1	muensterschezeitung.de	1
bassirat.net	1	mum.edu	1
batesstudent.com	1	munster-express.ie	1
bbc.co.uk	83	mural.com	1
bbv-net.de	1	murraystate.net	1
bdnews24.com	1	mutareradio.blogspot.com	1
bea.gov	1	myjoyonline.com	1
beds.ac.uk	1	mynaijanews.com	1
belizetimes.bz	1	mysanantonio.com	3
bellinghamherald.com	1	mysinchew.com	1
bentley.edu	1	mytechlaw.law.ttu.edu	1
berkeley.edu	5	mywesttexas.com	3
berlingske.dk	1	mz-web.de	5
berlinonline.de	1	mzz.gov.si	1
bgnewsnet.com	1	na.se	1
bhutanobserver.com	1	nachrichten.at	1
bienpublic.com	1	nactem.ac.uk	1
billboard.com	2	naftemporiki.gr	1
billingsgazette.net	1	napi.hu	1
binghamton.edu	3	naplesnews.com	2
biofuelreview.com	1	registerblog.splinder.com	1
bizhallmark.com	1	narinjara.com	1
biznespolska.pl	1	nation.co.ke	1
blackboard.uic.edu	1	nation.com.pk	1
blackbookmag.com	2	nation.ittefaq.com	1
bleskove.aktualne.centrum.cz	1	nationalarchives.gov.uk	1
blic.rs	1	NationalEnquirer.com	1
blikk.hu	2	nationalgeographic.com	1
blikopnieuws.nl	1	nationalpointonline.com	1
blogsagainsthillary.com	1	nationalpost.com	6
blt.se	1	nationalreview.com	1
bna.bh	1	nature.com	12
bnamericas.com	1	navegalo.com	1
bndestem.nl	1	navy.mil	1
bns.lt	1	ncarts.edu	1
bohemia.cubaweb.cu	1	ncca.com	4
boingboing.net	1	nczas.com	1
borastidning.se	1	nd.nl	1
bordermail.com.au	1	ndtv.com	6
borsen.dk	1	nemzetihirhalo.hu	1
boston.com	4	nepalbc.com	1
bostonherald.com	2	nepalnews.com	1
brabantsdagblad.nl	1	nepszava.hu	3
bradenton.com	5	network.nationalpost.com	1
brainsnap.com	1	netzeitung.de	4

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
brandeis.edu	2	neues-deutschland.de	1
breakingnews.ie	2	nevadaappeal.com	2
brightsurf.com	1	neviditelnypes.lidovky.cz	1
bris.ac.uk	2	newamerica.net	4
bristol.ac.uk	2	news-journalonline.com	2
bt.dk/	1	news-register.net	1
bu.edu	1	news-reporter.com	1
budapester.hu	1	news.com	3
budapestsun.com	1	news.com.au	24
buenosairesherald.com	1	news.com.au/couriermail/	5
buffalo.edu	1	news.com.au/dailytelegraph	5
buffalonews.com	1	news.com.au/heraldsun	6
buffalostate.edu	1	news.com.com	1
bugandapost.com	1	news.dailytrust.com	1
bulgaria-weekly.com	1	news.lk	1
bulgariagazette.com	1	news24.com	1
business.theage.com.au	1	newsbusters.org	1
businessdayonline.com	1	newscientist.com	1
businessfinancemag.com	1	newsclick.de	2
businessinafrica.net	1	newsday.com	13
businessnews-bd.com	1	newsevents.tcu.edu	2
businessonline.it	1	newsgator.com	1
businessweek.com	4	newsmax.com	8
businessweekly.co.zw	1	newsobserver.com	1
butler.edu	2	newsoftheworld.co.uk	3
buzzmachine.com	1	newsok.com	2
buzznet.com	1	newsonjapan.com	1
buzzpatrol.com	1	newstin.com	1
bydgoski.pl	1	newsweek.com	10
byu.edu	1	newtimes.co.rw	1
byuh.edu	2	newvision.co.ug	1
cadenaser.com	1	newz.ro	1
campusapps.fullerton.edu	1	ngz-online.de	6
canada.com	1	nicaraguanpost.com	1
canada.com	6	nieuwsblad.be	1
canada.com	1	nigerianewsdirect.com	1
canada.com	1	nih.gov	2
canada.com	1	ninemsn.com.au	1
canada.com	1	nj.com	1
canada.com	1	nkomazivoice.com	1
canada.com	1	nme.com	1
canada.com	1	nmt.edu	1
canada.com	1	nohillaryclinton.com	1
canadaeast.com	5	nol.hu	1
capebretonpost.com	1	nola.com	3
capecodonline.com	4	noordhollandsdagblad.nl	1
capital.ro	1	nordjyske.dk	1
capitol-college.edu	1	norran.se	1
captainsquartersblog.com	1	northampton.ac.uk	1
care.org	1	northernstar.info	1
cas.sk	1	northjersey.com	1
casafree.com	1	novaguarda.pt	1
case.edu	2	novayagazeta.ru	1
catholicnewsagency.com	1	noveslovo.sk	1
cbc.ca	1	novinite.com	1
cbn.co.za	1	novonews.lv	1
cbs58.com	1	nowmagazine.co.uk	3
cbs7kosa.com	1	npr.org	13
cbsnews.com	26	nra.lv	1
cdt.org	1	nrc.nl	1
celebitchy.com	1	nsd.se	1
celebritat.blogspot.com	1	nsf.gov	2
celebrity-gossip.net	1	nsula.edu	1
celebrity-sensuo.us	1	nt.se	1
celebritydresses.blogspot.com	1	nto.pl	1
celebrityfashionparty.blogspot.com	1	number-10.gov.uk	1
celebritynewslive.blogspot.com	1	nwt.se	1
celebrityscum.com	1	nwzonline.de	3
celebrityspotlight.co.uk	1	nyan.aland.fi	1
ceskenoviny.cz	2	nyasatimes.com	1
channel4.com	1	nydailynews.com	17
channelnewsasia.com	6	nymag.com	4
chareidio.com	1	nypost.com	18
charleston.net	2	nytimes.com	16
chattershmatter.com	1	nzherald.co.nz	3
chicagopublicradio.org	2	observer.gm	1
chicagotribune.com	8	observer.ug	1
chinadaily.com.cn	6	ocregister.com	3
chinapost.com.tw	2	odia.terra.com.br	5
chinaview.cn	1	oem.com.mx	2
chiosnews.com	1	ofigueirense.com	1
christianitytoday.com	1	ogaden.com	1
christiantoday.com	3	oglobo.globo.com	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
chron.com	5	ohio.com	6
chronicle.com	1	ohtuleht.ee	1
chronicle.uchicago.edu	1	ojo.com.pe	1
churchtimes.co.uk	1	ojogo.pt	1
cincinnati.com	3	oneindia.in	4
cinemablend.com	2	onet.pl	1
citizen.co.za	2	op-for.com	1
city.corriere.it/	1	op-marburg.de	5
citypaper.ee	1	op-online.de	5
civil.ge	1	op.se	2
cjr.org	1	open.ac.uk	1
clarin.com	1	oregional.pt	1
clicanoo.com	1	orf.at	1
clickz.com	1	ori.hhs.gov	1
clintonherald.com	3	orlandosentinel.com	2
clintonnc.com	2	osoblog.tv	1
clippertoday.com	1	ostran.se	1
clujeanul.ro	1	ostsee-zeitung.de	1
cmu.edu	1	ottawacitizen.com	1
cnews.canoe.ca	5	ourmidland.com	2
cnn.com	27	overthelimit.info	1
cnycentral.com	4	overthetop.beloblog.com	1
collegian.com	1	ovimagazine.com	1
colombopage.com	1	ox.ac.uk	1
colorado.edu	1	oxfam.org	1
coloradoan.com	3	oxfordjournals.org	188
columbia.edu	2	pacificfreepress.com	1
columbustech.edu	1	pacificnewscenter.com	1
communication.go.ke	1	palinforvp.blogspot.com	1
compassnews.net	1	palmbeachpost.com	2
computer.org	24	palomar.it	1
congoplanet.com	2	panama-guide.com	1
connexionfrance.com	1	panarmenian.net	1
conservatives.com	1	panorama.am	1
cornell.edu	1	parkiet.com	1
corner.nationalreview.com	1	parkvideo.net	1
correio.editpress.lu	1	parool.nl	1
correiodominho.com	1	pathfinder.gr	1
correiomanha.pt	1	pb.pl	1
corren.se	1	pbs.org	2
correoperu.com.pe	1	pcworld.com	1
corriere.it	6	people.com	3
cosmo.gr	4	people.com.cn	4
cosmosmagazine.com	6	peoples-view.org	1
cotidianul.ro	1	peoplesdaily-online.com	1
courant.com	4	perezhilton.com	1
courier-journal.com	1	periodistadigital.com	1
courrierinternational.com	1	peru21.pe	1
covenant.edu	1	peruviantimes.com	1
cphpost.dk	1	philelefttheros.com	4
cpilive.net	2	philly.com	7
cretegazette.com	1	philstar.com	2
critica.com.pa	1	phnompenhpost.com	1
croatiantimes.com	1	physicstoday.org	1
cronica.com.ec	1	physicstoday.org	1
cronica.com.mx	4	pietarsaarensanomat.fi	1
cronica.com.mx	2	pitt.edu	1
cronista.com	4	pittnews.com	1
crooksandliars.com	1	planetapesca.com	1
csmonitor.com	6	playacommunity.com	1
cstx.gov	1	plosone.org	40
ctv.ca	1	pnas.org	3
cuny.edu	2	pohjalainen.fi	1
dagen.se	3	pohjolansanomat.fi	1
daily.stanford.edu	1	poligazette.com	1
dailycollegian.com	1	polishmarket.com	1
dailyexpress.co.uk	2	polishnews.com	1
dailykos.com	1	politico.com	2
dailymail.co.uk	2	politiken.dk	1
dailymirror.lk	1	polskieradio.pl	1
dailypennsylvanian.com	1	pomorska.pl	1
dailyprincetonian.com	1	poponut.com	1
dailyrecord.co.uk	3	popsci.com	7
dailyreportonline.com	1	popsugar.com	1
dailystar.co.uk	2	popularmechanics.com	4
dailystar.com.lb	1	popwatch.ew.com	1
dailystaregypt.com	1	port.ac.uk	3
dailytexanonline.com	1	portafolio.com.co	1
dailytimes.com	1	portal.tt.com	1
dailytimes.com.pk	6	portfolio.com	1
dailytrojan.com	1	portugalresident.com	3
dalademokraten.se	1	post-gazette.com	16
dallasnews.com	6	postimees.ee	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
DarmstadtNews.de	1	postnewsline.com	1
davenport.edu	1	pottsmerc.com	4
dawn.com	1	povoasemanario.pt	1
dbtechno.com	1	powerlineblog.com	1
dcwblogs.com	1	pr-inside.com	4
debka.com	1	prague-tribune.cz	1
defenselink.mil	5	praguemonitor.com	1
defimedia.info	1	pravda.sk	1
del.icio.us	1	prensa-latina.cu	1
delawareonline.com	3	prensalibre.com	1
delmagyar.hu	1	press-gr.blogspot.com/	1
delo.si	1	pressedelanation.com	1
demerarawaves.com/	1	presstv.ir	11
democrats.org	1	primorske.si	1
demokrata.hu	1	princeton.edu	1
demorgen.be	1	prnewswire.com	2
denik.cz	1	proceso.hn	1
denmark.dk	1	profit.bg	1
denverpost.com	3	prospect.org	1
depechedekabylie.com	1	publicbroadcasting.net	1
depers.nl	1	publicrelations.uncc.edu	1
derby.ac.uk	1	purdue.edu	1
derStandard.at	1	pzc.nl	1
derstandarddigital.at	1	qctimes.com	8
derwesten.de	3	qn.quotidiano.net	5
deseretnews.com	1	quickdfw.com	1
destak.pt	1	radio.cz	3
desu.edu	1	radionetherlands.nl	1
detnews.com	9	radionz.co.nz	1
df.cl	1	radiovaticana.org	2
dhnet.be	1	raeng.org.uk	1
di.se	1	rai.it	1
diaadia.com.pa	1	rainews24.it	5
diariocoimbra.pt	1	rave.ac.uk	1
diariocolatino.com	1	reconquista.pt	1
diariodelhuila.com	1	recordnet.com	5
diariodosacores.pt	1	redherring.com	2
diarioelpueblo.com.uy	1	redorbit.com	4
diariolasamericas.com	9	redpepper.org.uk	1
diariopopular.com.br	1	refdag.nl	1
diariovasco.com	1	reforma.com	1
didimtoday.com	1	regards.fr	1
diena.lt	1	regiaodeleiria.pt	1
diena.lv	1	rep-am.com	2
diepresse.com	1	reporter.gr	1
digitalhit.com	1	repubblica.it	1
digitalspy.co.uk	1	republikein.com	1
dimokratiki.gr	1	reuters.com	21
dimokratiki.org	1	rewardsforjustice.net	1
diplomatie.gouv.fr	1	rferl.org	1
dir.salon.com	1	rfi.fr	1
discovermagazine.com	9	rhein-zeitung.de	1
dispatch.co.za	1	rian.ru	1
dispatch.com	1	rice.edu	2
dissidentvoice.org	1	ricethresher.org	4
dlisted.com	1	rightontheright.com	1
dn.se	1	ripon.edu	1
dna.fr	2	rit.edu	2
dnaindia.com	1	rizospastis.gr	1
dnevnik.bg	1	roanoke.com	5
dnevnik.si	1	rochester.edu	1
dnn-online.de	1	rockefeller.edu	1
dod.gov	4	rockymountainnews.com	1
Dolezite.sk	1	romanialibera.com	1
domainabc.hu	1	romanialibera.ro	1
dose.ca	1	roosevelt.edu	1
dosmundos.com	2	rote-hilfe.de	1
douglascountysentinel.com	1	rp-online.de	1
dr.dk	1	rp.pl	1
dresden-news.com	1	rpi.edu	1
dsc.discovery.com	4	rte.ie	2
dsl.psu.edu	1	ruhrnachrichten.de	1
dt.se	1	rumorficial.com	1
duke.edu	6	russiatoday.ru	1
dutchnews.nl	1	rutgers.edu	1
dvhn.nl	1	sabcnews.com	7
dw-world.de	5	sac.ac.uk	1
dziennik.pl	1	sacbee.com	1
dziennikwschodni.pl	1	sagoodnews.co.za	1
e-bangladesh.org	1	saharareporters.com/index.php	1
e-grammes.gr	1	salem-news.com	1
e-tipos.com	1	salford.ac.uk	1
ea.com.py	1	salzburg.com	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
earlham.edu	1	santafetimes.com	1
earthtimes.org	2	santiagotimes.cl	1
eastandard.net	2	sapo.pt	1
ec.europa.eu/malta-mt	1	sapo.pt	1
echoroukonline.com	1	sarah.swbts.edu	1
economico.sapo.pt	1	satakunnankansa.fi	1
economist.com	10	savonsanomat.fi	1
economist.com.na	1	sbpost.ie	1
ed.nl	1	sbs.com.au	1
edennewspaper.com	1	schuttevaer.nl	1
edestad.nl	1	sciam.com	12
edgewood.edu	1	scidev.net	1
eeo.com.cn	1	sciencedaily.com	9
eestikirik.ee	1	sciencemag.org	1
efluxmedia.com	5	scitation.aip.org	11
egossip.com	1	scmp.com	7
eitb24.com	1	scotlandonsunday.com	1
ej.iop.org	47	scotsman.com	5
ekipnews.com	1	scu.edu	1
ekspress.ee	1	sdsu.edu	1
ekstrabladet.dk	1	se.pl	1
elargentino.com	7	seacoastonline.com	1
elcanillita.com	2	seattlepi.nwsource.com	14
elcolombiano.com	2	seattletimes.nwsource.com	13
elcomercio.com	1	segabg.com	1
elcomercio.pe	1	semanario.ucr.ac.cr	1
elcorreodigital.com	1	senegambianews.com	1
eldiario.com.ec	1	sermitsiaq.gl	1
eldiariodechihuahua.com.mx	1	servitoros.gr	1
eldiariony.com	1	sethgodin.typepad.com	1
elespectador.com	6	sfgate.com	3
elfinanciero.com.mx	1	shanghaidaily.com	3
ellibertador.hn	1	shc.edu	1
elliderusa.com	7	shef.ac.uk	1
elmercurio.com.ec	1	shortnews.com	1
elmigrante.com.ec	1	showbiznews.info	1
elmiradorparaguayo.com	1	showbizspy.com	1
elmoudjahid.com	1	shz.de	1
elmundo.es	1	sierraexpressmedia.com	1
elnacionaltarija.com	1	sierramaestra.cu	1
elnorte.com	1	sify.com	2
elnuevodia.com.co	1	sigmalive.com	1
elnuevodiario.com.ni	1	sigmalive.com/simerini	1
elnuevoherald.com	5	signonsandiego.com	2
elpais.com	4	silkroadintelligencer.com	1
elpais.com.co	1	siouxcityjournal.com	8
elpais.com.uy	1	skai.gr	2
elperiodico.com.gt	1	skrastas.lt	1
elporvenir.com.mx	1	sktoday.com	1
elsalvador.com	1	sky.com	4
elsiglo.cl	1	skylife.it	1
eltabloide.com.co	1	slashdot.org	1
eluniversal.com	4	slate.com	3
eluniversal.com.co	1	smh.com.au	6
eluniversal.com.mx	1	smp.se	1
eluniverso.com	1	smudailycampus.com	1
elvoceromi.com	5	sn.se	1
elwatan.com	1	snc.edu	1
emarrakech.info	1	socialistworker.org	1
emerson.edu	1	soester-anzeiger.de	2
en.rian.ru	1	sofiaecho.com	1
enet.gr	1	sol.de	2
english.pravda.ru	1	soleno-morsko.com	1
engr.uiuc.edu	1	somalipress.com	1
enidnews.com	1	sondagsavisen.dk	1
ennaharonline.com	1	soton.ac.uk	1
enquirer.com	1	southbendtribune.com	1
entertainmentandshowbiz.com	1	sowetan.co.za	2
entertainmentavenue.com	1	spectator.co.uk	4
entertainmentwise.com	1	spectator.org	1
eonline.com	5	spectator.sk	1
epl.ee	1	spiegel.de	6
epsrc.ac.uk	1	spokesmanreview.com	1
epsy.tamu.edu	1	sportsister.com	1
erhvervsbladet.dk	1	spreconi.it	1
ert.gr	2	sptimes.ru	1
esa.int	1	srilankaguardian.org	1
esaimaa.fi	1	st-andrews.ac.uk	1
esf.edu	1	st.nu	1
espressonews.gr	1	sta.si	5
ess.fi	1	stabroeknews.com	1
estadao.com.br	1	standaard.be	2
esteri.it	1	standartnews.com	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
estrategia.cl	1	stanford.edu	1
estrepublicain.fr	1	stanford.edu	1
etonline.com	1	stanford.org	1
etrend.sk	1	star-telegram.com	1
eunyheter.se	1	star.com.jo	1
euobserver.com	3	starbulletin.com	4
eurasianet.org	1	starmagazine.com	1
eurekalert.org	7	starmuscle.com	1
euronews.net	6	starpulse.com	2
europapress.es	1	startribune.com	8
event.iastate.edu	1	starvalleyindependent.com	1
events.utah.edu	1	state.gov	11
events3.stanford.edu	1	statenews.com	1
evz.ro	1	statesman.com	2
ew.com	2	stephens.edu	1
examiner.com	3	stern.de	5
exonline.com.mx	2	stfranciscollege.edu	1
expreso.ec	1	stiften.dk	1
express.co.uk	4	stltoday.com	2
express.de	1	stmartin.edu	1
expressen.se	1	stoxos.gr	1
expressindia.com	1	stuff.co.nz	13
extra.ec	1	stuttgarter-zeitung.de	1
extremetech.com	2	sudanvisiondaily.com	1
falmouth.ac.uk	1	sudonline.sn	1
famagusta-gazette.com	1	sudouest.com	1
fanbolt.com	1	sueddeutsche.de	7
faz.net	6	suedkurier.de	1
fd.nl	1	suedwest-aktiv.de	1
fds.duke.edu	1	sundaymail.co.uk	2
feedburner.com	13	sundaystandard.info	1
feedmegossip.com	1	sundaysun.co.uk	1
fema.gov	1	sundaytimes.lk	1
femalefirst.co.uk	2	sunherald.com	1
feverishthoughts.com	1	suntimes.com	1
fhsu.edu	1	surrey.ac.uk	1
fimotro.blogspot.com	1	sva.edu	1
financialexpress.com	1	svd.se	1
financialmirror.com	1	svz.de	1
fjgirls.com	1	swau.edu	1
fn.hu/hetilap	1	swazilive.com	1
focus.de	6	sweden.gov.se	1
folkbladet.nu	1	swissinfo.ch	4
forbes.com	12	sydin.fi	1
foroyaa.gm	2	sydsvenskan.se	1
foxnews.com	12	syndication.boston.com	5
fr-online.de	5	sz-online.de	4
france2.fr	7	SZOn.de	1
france24.com	11	tageblatt.de	1
francesoir.fr	8	tageblatt.editpress.lu	1
freace.de	1	tagesspiegel.de	5
fredonia.edu	1	talcualdigital.com	1
freep.com	2	talkingpointsmemo.com	2
frieschdagblad.nl	1	talkzimbabwe.com	1
ft.com	19	taloussanomat.fi	1
ftd.de	1	tamu.edu	1
fuldaerzeitung.de	5	tanea.gr	1
fyens.dk	1	targetnewsonline.com	1
fynsamtsavis.dk	1	tbo.com	3
galwayindependent.com	1	tcf.org	1
gambianow.com	1	tctubantia.nl	1
gardianul.ro	1	teamxbox.com	1
gawker.com	5	techdirt.com	1
GayWired.com	1	technorati.com	2
gazetalubuska.pl	1	techpresident.com	1
gazetaolsztynska.wm.pl	1	techreview.com	3
gazetaprawna.pl	1	tees.ac.uk	3
gazettebw.com	1	teesdalemercury.co.uk	1
gazzettadiparma.it	1	tehrantimes.com	1
gbr.pepperdine.edu	1	telegraaf.nl	1
gd.se	1	telegrafo.com.ec	1
gds.ro	1	telegraph.co.uk	5
gelderlander.nl	1	telegraphindia.com	3
gelocal.it	1	teletext.co.uk	6
general-anzeiger-bonn.de	1	terra-economica.info	1
georgetowncollege.edu	1	texasmonthly.com	1
georgiasouthern.edu	1	texelsecourant.nl	1
gic.nl	1	tgcom.mediaset.it	4
giron.co.cu	1	thaindian.com	1
globalresearch.ca	1	theage.com.au	4
globegazette.com	1	theaustralian.news.com.au	3
goettinger-tageblatt.de	6	theblemish.com	1
goldcoast.com.au	1	theboquetetimes.com	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
gomemphis.com	3	thebulgarianpost.com	1
gooieneemlander.nl	1	thecanadianpress.com	3
gop.com	1	thecelebritycafe.com	2
gopconvention2008.com	1	thechanticleeronline.com	1
gossipboulevard.com	1	thecheers.org	6
goucher.edu	1	thecolombotimes.com	1
gov.bw	1	thecolumbiastar.com	1
gozonews.com	1	theconservativevoice.com	1
gp.se	1	thecostaricanews.com	1
grandbaie.mu	1	thecurrent.theatlantic.com	1
granitefallsnews.com	1	thedailystar.net	1
graphic.pepperdine.edu	1	theday.com	1
greatindaba.com	1	thedidymian.com	1
greatzimbabwenews.com	1	thefinancialexpress-bd.com/	1
greeleytrib.com	2	thefrontierpost.com	3
greenleft.org.au	1	theglobeandmail.com	8
greenpeace.org	1	theguardian.pe.ca	2
grist.org	2	thehawaiichannel.com	1
grnet.it	1	theherald.co.uk	2
groene.nl	1	thehimalayantimes.com	1
gs24.pl	1	thehindu.com	1
guampdn.com	1	thehollywoodgossip.com	1
guardian.bz	1	thehoya.com	1
guardian.co.uk	20	thehurricaneonline.com	1
guardian.co.uk	2	theindependent.co.zw	1
guatemala-times.com	1	theindependent.com	1
gulf-daily-news.com	2	theintelligencer.com	1
gustavus.edu	2	thelocal.se	1
guyanalive.com	1	themedialine.org	1
guyananewstoday.com	1	themoscowtimes.com	1
guyanaobservernews.org	1	thenation.com	1
gva.be	1	thenews-gazette.com	1
haaretz.com	5	thenews.com.pk	4
haarlemsdagblad.nl	1	thenewsng.com	1
haitipress.net63.net	1	thenewstribune.com	4
halifax.metronews.ca	7	theolivepress.es	1
hallandsposten.se	1	theolympian.com	1
halonoviny.cz	1	thepoint.gm	1
hameensanomat.fi	1	theprairiestar.com	1
handelsblatt.com	1	therant.us	1
hararetribune.com	1	theregister.co.uk	2
harper-adams.ac.uk	2	thestandard.com.hk	2
harvard.edu	1	thestar.com	3
harvard.edu	1	thestar.com.my	3
harvard.edu	1	thesun.co.uk	1
harvard.edu	5	thetandd.com	3
harvard.edu	1	thetimes-tribune.com	1
haveeru.com.mv	1	thetimesofnigeria.com	1
hawaiinews.com	1	thevoicebw.com	1
hawaiitribune-herald.com	1	thewhig.com	1
haz.de	4	thezimbabwean.co.uk	1
hbindependent.com	5	thezimbabwetimes.com	1
hbvl.be	1	thinkprogress.org	1
hd.se	1	thinkspain.com	1
healthnewsdigest.com	1	thisdayonline.com	1
heat.co.za	1	thisis.co.uk	4
hefce.ac.uk	3	thisisbath.co.uk	1
hellomagazine.com	1	thisisbristol.co.uk	1
helsinkitimes.fi	1	thisiscornwall.co.uk	1
heraldsun.com	3	thisisexeter.co.uk	1
hfxnews.ca	2	thisisgloucestershire.co.uk	1
hiiraan.com	1	thisishull.com	1
hillaryclinton.com	3	thisiskent.co.uk	1
hillaryis44.org	2	thisisleicestershire.co.uk	1
hindustantimes.com	3	ticiero.com	1
hln.be	1	ticotimes.net	1
hn.se	1	tijd.be	1
hnonline.sk	1	time.com	9
hollyscoop.com	1	timeslive.co.za	1
hollywood.com	1	timesofmalta.com	1
hollywoodrag.com	1	timesonline.co.uk	7
hollywoodreporter.com	2	timesonline.typepad.com	2
holymoly.co.uk	1	tlz.de	2
hondurasnews.tv	1	tmz.com	1
hondurasthisweek.com	1	tnr.com	2
honoluluadvertiser.com	5	tntech.edu	1
hope.ac.uk	1	to.com.pl	1
hotair.com	1	todayszaman.com	2
howardshome.com	2	togonews.canalblog.com	1
hoy.com.ec	1	togosite.com	1
hoycanelones.com.uy	1	tol.cz	1
hoyinternet.com	3	topix.com	1
hpu.edu	1	topix.com	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
hrw.org	1	topix.net	5
hs.fi	1	tradearabia.com	4
hsc.edu	1	transworldnews.com	3
hsph.harvard.edu	1	treehugger.com	1
http://polskatimes.pl	1	trevecca.edu	1
http://www.bosnia.org.uk/news	1	tribstar.com	1
hub.ou.edu	1	tribune.ie	1
hudson.org	1	trinidadexpress.com	1
huffingtonpost.com	4	trinitybiblecollege.edu	1
humanite.fr	4	trinitycollege.edu	1
hurriyetdailynews.com	1	troktiko.blogspot.com	1
hw.ac.uk	2	trouw.nl	1
i-newswire.com	1	tufts.edu	1
iaea.org	1	Tundzha.Info	1
iafrica.com	1	tunisiadaily.com	1
ibiza-spotlight.com	1	turkishpress.com	1
ibox.bg	1	turkishweekly.net	10
icelandreview.com	1	turunsanomat.fi	1
icydk.com	1	tvguide.com	2
idnes.cz	1	tvxs.gr	1
ihned.cz	1	twitter.com	1
iht.com	3	twitter.com/BarackObama	1
ilgiornale.it	1	twitter.com/hillaryclinton	1
ilkka.fi	1	typicallyspanish.com	1
ilmessaggero.it	1	typos.com.cy	1
ilsole24ore.com	3	ua.edu	1
iltalehti.fi	1	uchicago.edu	1
iltasanomat.fi	1	uchicago.edu	1
imerazante.gr	1	uchospitals.edu	1
imerisia.gr	1	ucl.ac.uk	3
imnotobsessed.com	1	ucla.edu	3
imperial.ac.uk	1	ucla.edu	1
in-forum.com	1	ucmo.edu	1
in.gr	2	uconn.edu	1
independent-bangladesh.com	6	ucop.edu	1
independent.co.ug	1	ucsc.edu	1
independent.co.uk	13	ucsf.edu	3
independent.ie	3	ucsfhealth.org	1
indianexpress.com	1	udel.edu	1
indiatimes.com	5	ugee.com	1
inform.kz	1	ujszo.com	1
informador.com.mx	1	ukraine-observer.com	1
information.dk	1	ulster.ac.uk	4
infoworld.com	1	ultimahora.com	1
infoworld.com	1	ultimasnoticiasdiario.com	1
ing.dk	1	ultraclear.net	1
inquirer.net	7	umaine.edu	1
inrich.com	2	umassd.edu	1
insidesomalia.org	1	umcrookston.edu	2
instapundit.com	1	umd.edu	1
intermediair.nl	1	umich.edu	1
inthenews.co.uk	7	umich.edu	1
invasor.cu	1	umich.edu	1
investors.com	2	umkc.edu	1
inyourpocket.com	1	umn.edu	1
iol.co.za	3	umr.edu	1
iran-daily.com	1	unian.net	1
ireland.com	9	unionesarda.it	3
irishexaminer.com	3	universetoday.com	1
irishtimes.com	1	universitas.uio.no	1
islamische-zeitung.de	1	universityofcalifornia.edu	1
israelhaiom.com	1	unmc.edu	1
israelnationalnews.com	1	unr.edu	1
isthmian.net	1	unt.edu	1
it.iu.edu	1	unt.se	1
itn.co.uk	2	up.edu	1
itp.net	3	upenn.edu	3
j-bradford-delong.net	1	upi.com	5
jacksonville.com	4	ursuline.edu	1
jamaica-star.com	1	uruknet.de	1
jamaicagleaner.com	1	usatoday.com	15
jamaicanewsbulletin.com	1	usc.edu	2
jamaicaobserver.com	1	usf.edu	1
jamesfallows.theatlantic.com	1	usi.edu	1
jamestownpress.com	1	usmagazine.com	1
japantimes.co.jp	2	usnews.com	2
japantoday.com	1	utah.edu	2
javno.com	1	utdallas.edu	1
jbonline.terra.com.br	2	utexas.edu	1
jfb.hu	1	utica.edu	1
jhu.edu	1	utk.edu	1
jimmatimes.com	3	utsouthwestern.edu	1
jmlr.csail.mit.edu	1	uusimaa.fi	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
jnytt.se	1	uvi.gov.si	1
jobbik.com	1	uwf.edu	1
johnmccain.com	3	uwsuper.edu	3
jonesbahamas.com	1	uwyo.edu	1
jornaldenegocios.pt	1	uzbekistannews.net	1
jornaldiario.com	1	valdosta.edu	1
jornaldofundao.pt	1	vanderbilt.edu	1
journalperu.com	1	vanguardia.co.cu	1
journalstar.com	4	vanguardia.com.mx	1
jp.dk	1	vanguardngr.com	1
jpost.com	6	variety.com	1
jsonline.com	8	vasabladet.fi	1
jta.org	6	vaterland.li	1
jungewelt.de	4	ve.lt	1
jurnalul.ro	1	vecer.com	1
justjared.buzznet.com	1	vejleamtsfolkeblad.dk	1
jv.dk	1	venezuelanalysis.com	1
ka-news.de	3	verkkouutiset.fi	1
ka-set.info	1	vestnesis.lv	1
kabulpress.org	1	vf.se	1
kainuunsanomat.fi	1	vg.hu	1
kaladanpress.org	1	vibeghana.com	1
kaleo.org	6	villagevoice.com	1
kansascity.com	1	vindy.com	1
kansascity.com	8	vipglamour.net	1
kapitals.lv	1	virgilio.it	5
karjalainen.fi	1	virginia.edu	2
kataweb.it	2	virginia.edu	1
katestone.wordpress.com	1	virginislandsdailynews.com	1
kathimerini.gr	6	vivelecanada.ca	1
katholieknieuwsblad.nl	1	vk.se	1
kathpress.at	1	vlt.se	1
kauppalehti.fi	1	vm.ee	1
kbtx.com	1	vmi.edu	1
kcra.com	2	vnunet.com	1
kerryman.ie	1	voanews.com	15
keskipohjanmaa.net	1	vob.org	1
keystone.edu	1	voicesnewspaper.com	1
kgmb9.com	1	vol.at	1
khaleejtimes.com	1	volksblatt.li	1
kingston.ac.uk	1	volksfreund.de	1
kingstonchronicle.wordpress.com	1	volkskrant.nl	1
kisalfold.hu	1	voltairenet.org	1
kitv.com	3	votehillary.org	2
kkuriren.se	1	votehope2008.com	1
klaipeda.diena.lt	1	votenader.org	1
kleinezeitung.at	1	vtnews.vt.edu	1
knoxnews.com	5	vz.lt	1
kokomotribune.com	1	wabash.edu	1
komentari.com	1	walesonline.co.uk	2
koreatimes.co.kr	1	walshcollege.edu	2
kotzot.com	1	warren-wilson.edu	1
kouvolansanomat.fi	1	washington.edu	1
krakowpost.com	1	washingtonbureau.typepad.com	1
krgv.com	1	washingtonmonthly.com	1
kristeligt-dagblad.dk	1	washingtonpost.com	15
kristianstadsbladet.se	1	washingtontimes.com	1
krone.at	1	watchblog.com	3
ksbw.com	5	wbz.com	2
ksl.com	2	wdetfm.org	1
ksta.de	1	weeklyblitz.net	1
ktbb.com	1	weeklystandard.com	1
ktvu.com	4	welt.de	7
kuenselonline.com	1	wenn.com	1
kullhadd.com	1	westerncourier.com	1
kurier.at	1	westga.edu	1
kurier.lt	1	westhawaiitoday.com	1
kurierlubelski.pl	1	westminster.edu	1
kuriren.nu	1	wfp.org	1
kuro5hin.org	1	whitehouse.gov	4
kutv.com	1	wicz.com	1
kxmb.com	8	wienerzeitung.at	1
kyivpost.com	1	wighsnews.se	1
kymensanomat.fi	1	wiley.com	90
kyodo.co.jp	5	willcoxrangenews.com	1
la-croix.com	5	winthrop.edu	2
la-razon.com	1	wired.com	4
la.com	1	wirtschaftsblatt.at	1
la7.it	1	wisc.edu	3
labour.org.uk	1	wit.edu	1
lacapital.com.ar	1	wiu.edu	1
lacuarta.cl	1	wksu.org	1
lademajagua.co.cu	1	wkuherald.com	1

Domain of Outlet	# Feeds	Domain of Outlet	# Feeds
ladepeche.fr	7	wlv.ac.uk	2
lafayette.edu	3	wlz-fz.de	4
lahora.com.ec	1	wmitchell.edu	1
laisvaslaikrastis.lt	1	wn.com	1
lakesunleader.com	1	wnba.com	1
lalibre.be	1	wnyc.org	3
lameuse.be	1	wolfram.com	1
lanacion.com.ar	1	womenshealthmag.com	1
lanacion.com.co	1	womentalksports.com	1
lanacion.com.py	1	wonkette.com	1
lankapolity.com	1	woofactor.com	1
lankasun.com	1	worldontheweb.com	4
lanouvellerepublique.fr	1	worldpress.org	2
lansi-savo.fi	1	wort.lu	1
lansivayla.fi	1	wprost.pl	1
laopinion.com	6	wsaz.com	1
laopinion.com.co	1	wsj.com	19
lapatria.com	1	wwaytv3.com	1
lapatriaenlinea.com	1	www.jnewswire.com	1
lapinkansa.fi	1	wyborcza.pl	1
laprensa.com.ar	1	wyomingnews.com	1
laprensa.com.ni	1	xpatloop.com	1
laprensagrafica.com	1	yahoo.com	14
laprovence.com	5	yeeeah.com	1
laraza.com	5	yementimes.com	1
larepublica.net	1	ylioppilaslehti.fi	1
larepublica.pe	1	ynetnews.com	1
lasegunda.com	1	yobserver.com	1
lastampa.it	1	ystadsallehanda.se	1
latech.edu	1	z.about.com	1
latercera.com	1	zadarbj.blogspot.com	1
latimes.com	15	zapernik.com	1
latribuna.hn	1	zeenews.com	6
latribune-online.com	1	zeit.de	5
latribune.fr	6	zenit.org	1
latviansonline.com	1	zerohora.clicrbs.com.br	1
lavanguardia.es	1	zf.ro	1
laverdad.com	1	zimbabwemetro.com	1
lavoixdunord.fr	1	zimbabwenewsonline.com	1
lavozdegalicia.es	1	zimbabweonlinepress.com	1
le-jeudi.editpress.lu	1	ziminternationalnews.com	1
le.ee	1	zimnewsroom.org	1
leadershipnigeria.com	1	zimnewswire.com	1
leadertelegram.com	2	ziua.ro	1
ledauphine.com	3	zougla.gr	1
leeuwardercourant.nl	1	zuidfriesland.nl	1
lefigaro.fr	6	zw.com.pl	1
leftword.blogdig.net	1	zz.lv	1

8.2 Supplementary Figures

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet title="XSL_formatting" type="text/xsl" href="/shared/bsp/xsl/rss/nolsol.xsl"?>

<rss xmlns:media="http://search.yahoo.com/mrss/" xmlns:atom="http://www.w3.org/2005/Atom" version="2.0">
  <channel>
    <title>BBC News - Home</title>
    <link>http://www.bbc.co.uk/go/rss/int/news/-/news/</link>
    <description>The latest stories from the Home section of the BBC News web site.</description>
    <language>en-gb</language>
    <lastBuildDate>Thu, 10 Mar 2011 16:29:07 GMT</lastBuildDate>
    <copyright>Copyright: (C) British Broadcasting Corporation, see http://news.bbc.co.uk/2/hi/help/rss/4498287.stm for terms and conditions of reuse.</copyright>
    <ttl>15</ttl>
    <atom:link href="http://feeds.bbci.co.uk/news/rss.xml" rel="self" type="application/rss+xml"/>
    <item>
      <title>Unions' anger over pensions plan</title>
      <description>Unions react angrily to a report proposing a radical overhaul of public sector pensions which would see millions working for longer.</description>
      <link>http://www.bbc.co.uk/go/rss/int/news/-/news/business-12687489</link>
      <guid isPermaLink="false">http://www.bbc.co.uk/news/business-12687489</guid>
      <pubDate>Thu, 10 Mar 2011 16:03:59 GMT</pubDate>
      <media:thumbnail width="66" height="49" url="http://news.bbcimg.co.uk/media/images/51600000/jpg/_51600580_011440451-1.jpg"/>
      <media:thumbnail width="144" height="81" url="http://news.bbcimg.co.uk/media/images/51600000/jpg/_51600578_011440451-1.jpg"/>
    </item>
    <item>
      ...
    </item>
  </channel>
</rss>
```

Figure 8.1: Example of the XML code of an news feed from BBC home page.

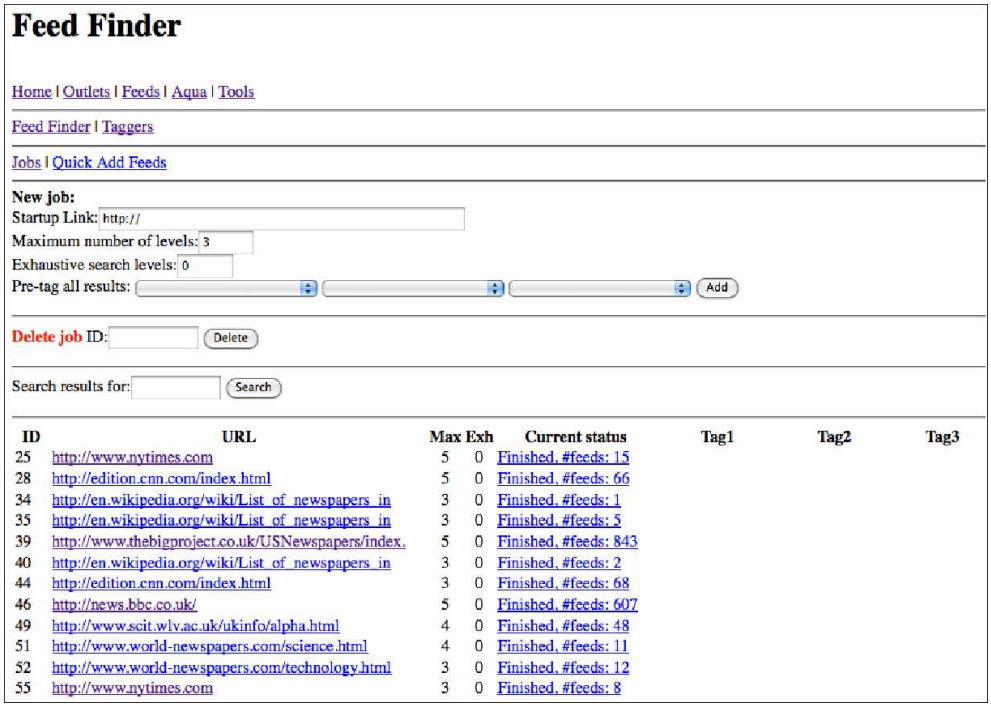


Figure 8.2: The front end of Feed Finder. User can add new crawl jobs and navigate the discovered feeds. The interface is part of SystemWatch, the front-end of NOAM system.

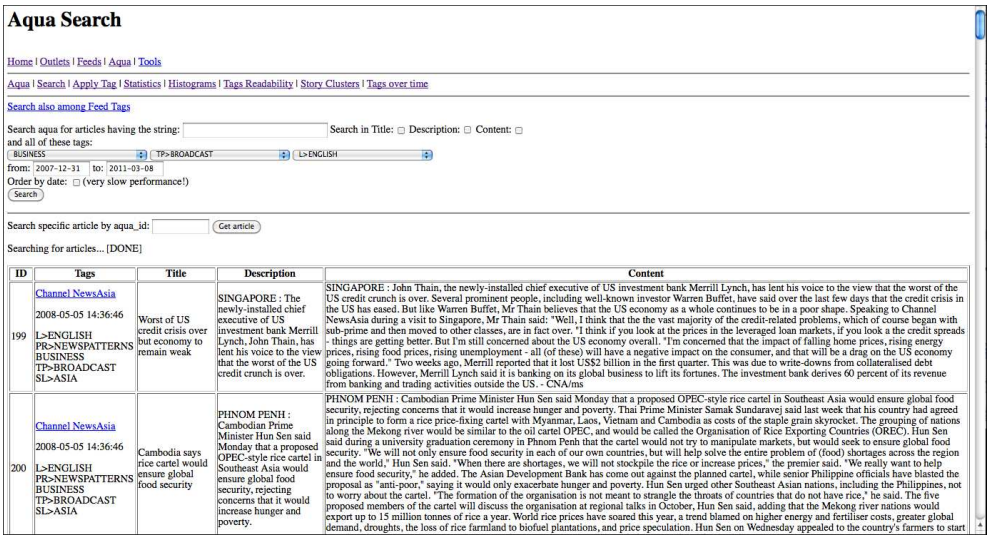


Figure 8.3: The Search page of SystemWatch provides the ability to search for articles that have a specific set of tags and contain specific keywords.

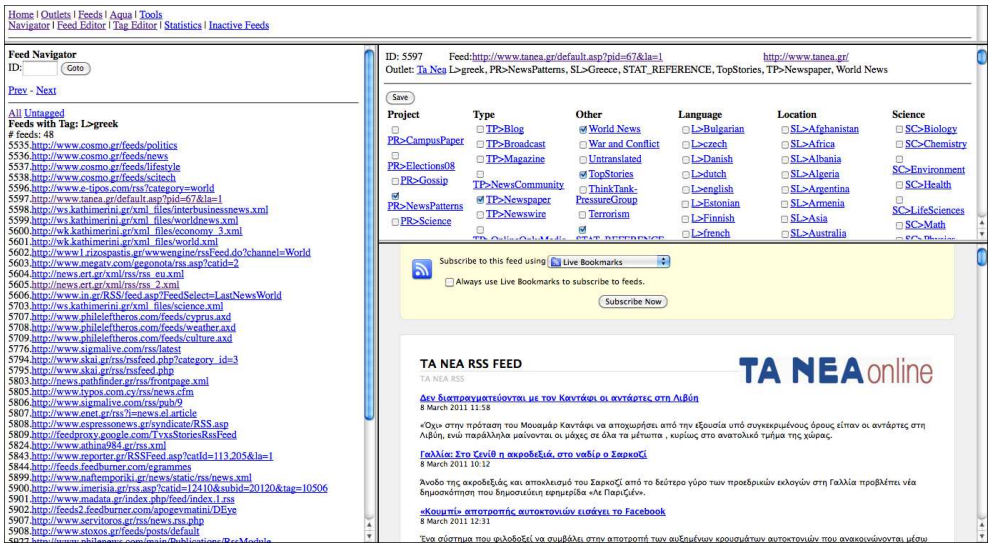


Figure 8.4: The Feed Navigator page allows the easy manual tagging of feeds. It also provides a preview of the feed and a list of subset of feeds for browsing among them.

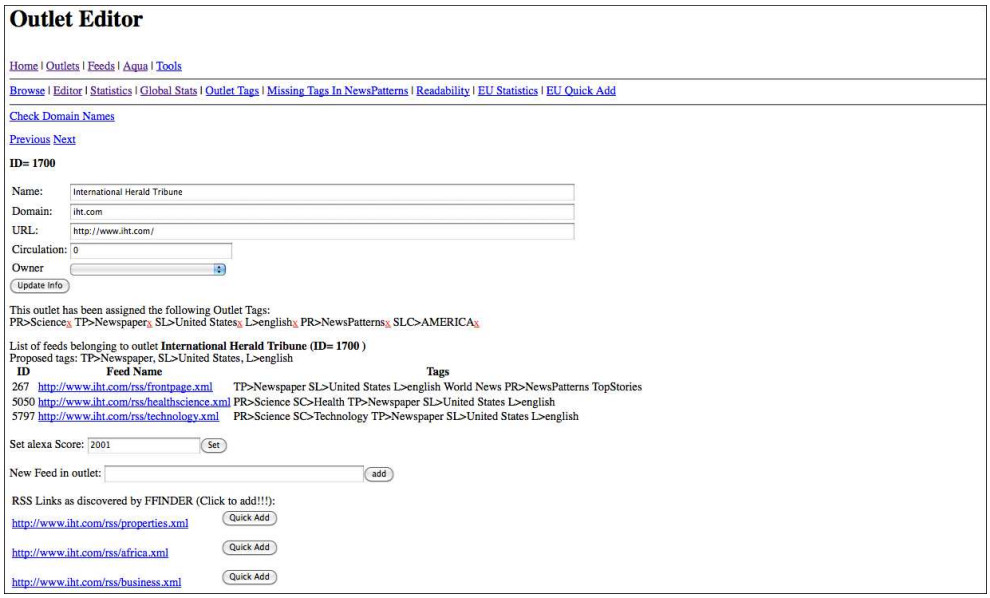


Figure 8.5: The Outlet Editor page allows the easy annotation of outets. It also provides a list of feeds, as found by Feed Finder module, that can be added to the outlet.

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under Curve
BRH	Best Reciprocal Hit
DB	Database
FIT	Found In Translation
GLM	Generalised Linear Model
IDF	Inverse Document Frequency
MDS	Multi-Dimensional Scaling
NOAM	News Outlets Analysis & Monitoring
PBSMT	Phrase Based Statistical Machine Translation
ROC	Received Operating Characteristic
RSS	Real Simple Syndication
SEM	Standard Error of the Mean
SMT	Statistical Machine Translation
SVM	Support Vector Machines
TF	Term Frequency -
VSM	Vector Space Model

Index

Bibliography

[1] L. Adamic and N. Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM New York, NY, USA, 2005.

[2] E. Adar and L. Adamic. Tracking information epidemics in blogspace. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 207–214, 2005.

[3] E. Adar, L. Zhang, L. Adamic, and R. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.

[4] S. Aday. Chasing the bad news: An analysis of 2005 Iraq and Afghanistan war coverage on NBC and Fox News Channel. *Journal of Communications*, 60:144–164, 2010.

[5] O. Ali and N. Cristianini. Information fusion for entity matching in unstructured data. In *Artificial Intelligence Applications and Innovations*, volume 339 of *IFIP Advances in Information and Communication Technology*, pages 162–169. Springer Boston, 2010.

[6] O. Ali, I. Flaounas, T. De Bie, and N. Cristianini. Celebrity Watch: Browsing news content by exploiting social intelligence. In *Machine Learning and Knowledge Discovery in Databases, European Conference (ECML PKDD)*, pages 613–616, 2011.

[7] O. Ali, I. Flaounas, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. Automating news content analysis: An application to gender bias and readability. In *Workshop on Applications of Pattern Analysis (WAPA)*, pages 36–43. JMLR: Workshop and Conference Proceedings, Windsor, UK, 2010.

[8] S. Allan. *Online news: Journalism and the Internet*. Open Univ Press, 2006.

[9] D. Ariely and G. Berns. Neuromarketing: the hope and hype of neuroimaging in business. *Nature Reviews Neuroscience*, 11:284–292, 2010.

[10] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May*, volume 25, pages 2200–2204, 2010.

[11] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *21st International Conference on Machine Learning*. Canada, 2004.

[12] W. Bainbridge. The scientific research potential of virtual worlds. *Science*, 317(5837):472, 2007.

[13] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.

[14] M. Bautin, C. Ward, A. Patil, and S. Skiena. Access: News and blog analysis for the social sciences. In *19th Int. World Wide Web Conference*, 2010.

[15] B.Boser, I. Guyon, and V.Vapnik. A training algorithm for optimal margin classifiers. In *5th Conference on Computational Learning Theory*, pages 144–152, 1992.

[16] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[17] S. Blackmore. *The Meme Machine*. Oxford University Press, 2000.

[18] T. Boyce, J. Kitzinger, and J. Lewis. *Science is everyday news - Review of UK Media Trends for the Office of Science and Innovation*. Cardiff School of Journalism, Media & Cultural Studies, Cardiff University, 2007.

[19] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, pages 462–465, 2005.

[20] P. Brown, S. Pietra, V. Pietra, and R. Mercer. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1994.

[21] P. Bruno, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, et al. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 53–58, 2006.

[22] P. Bruno and R. Steinberger. *Automatic Construction of Multilingual Name Dictionaries*. MIT Press - Advances in Neural Information Processing Systems Series (NIPS), 2009. 59–78 pp.

[23] P. Bruno, R. Steinberger, and C. Best. Automatic detection of quotations in multilingual news. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2007)*, pages 487–492, 2007.

[24] P. Bruno, H. Tanev, and M. Atkinson. Extracting and learning social networks out of multilingual news. In *Proceedings of the social networks and application tools workshop (SocNet-08)*, pages 13–16, 2008.

[25] H. Bunke. Error correcting graph matching: on the influence of the underlyingcost function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):917–922, 1999.

[26] C. Burges. *A Tutorial on Support Vector Machines for Pattern Recognition*. Kluwer Academic Publishers, Boston, USA, 1998.

[27] M. Burgoon, J. K. Burgoon, and M. Wilkinson. Writing style as a predictor of newspaper readership, satisfaction and image. *Journalism Quarterly*, 58:225–231, 1981.

[28] P. Butler. Visualizing friendships. http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919, Accessed January, 1st 2011, 2010.

[29] C. Cardie and J. Wilkerson. Text annotation for political science research. *Journal of Information Technology & Politics*, 5(1):1–6, 2008.

[30] W. Carlson and B. Thorne. *Applied statistical methods: for business, economics, and the social sciences*. Prentice Hall, 1997.

[31] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7):e11596, 2010.

[32] S. Chakrabari. *Mining the Web*. Elsevier Science, 2003.

[33] C. Chang and C. Lin. *LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[34] N. Chomsky. *Media control: The spectacular achievements of propaganda*. Seven Stories Press, 1997.

[35] C. Cioffi-Revilla. Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):259–271, 2010.

[36] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

[37] M. Cox and T. Cox. Multidimensional scaling. *Handbook of data visualization*, pages 315–347, 2008.

[38] K. Coyle. Mass digitization of books. *The Journal of Academic Librarianship*, 32:641–645, 2006.

[39] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based learning methods*. Cambridge University Press, 2000.

[40] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics*, pages 168–175. Philadelphia, USA, 2002.

[41] W. A. Danielson and S. D. Bryan. Readability of wire stories in eight news categories. *Journalism Quarterly*, 41:105–106, 1964.

[42] R. Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

[43] L. De Raedt. *Logical and Relational Learning*. Springer Verlag, 2008.

[44] S. Debnath, P. Mitra, N. Pal, and C. Giles. Automatic identification of informative sections of web pages. *IEEE transactions on knowledge and data engineering*, pages 1233–1246, 2005.

[45] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16:707–726, 2000.

[46] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representation for text categorization. In *7th International Conference on Information and Knowledge Management (CIKM)*, pages 148–155, 1998.

[47] S. Džeroski and N. Lavrač. *Relational data mining*. Springer Verlag, 2001.

[48] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[49] J. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences*, 101(40):14333–14337, 2004.

[50] Editorial. Defining the scientific method. *Nature Methods*, 6:237, 2009.

[51] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6(26):290–297, 1959.

[52] F. Esser. 'tabloidization' of news. *European Journal of Communication*, 14(3):291–324, 1999.

[53] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, pages 417–422, 2006.

[54] H. Evans. *Essential English for Journalists, Editors and Writers*. Pimlico, London, 2000.

[55] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[56] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[57] T. Fawcett. ROC graphs: Notes and practical considerations for researchers (Technical Report HPL-2003-4). *HP Laboratories, Palo Alto, CA, USA*, 2003.

[58] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[59] M. Feintuck and M. Varney. *Media Regulation, Public Interest and the Law*. Edinburgh University Press, 1999.

[60] M. Fernández and G. Valiente. A graph distance metric combining maximum common subgraph and minimum common supergraph. *Pattern Recognition Letters*, 22(6-7):753–758, 2001.

[61] P. Flach. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *International Conference on Machine Learning*, pages 194–201, 2003.

[62] I. Flaounas, O. Ali, M. Turchi, T. Snowsill, F. Nicart, T. De Bie, and N. Cristianini. NOAM: News Outlets Analysis and Monitoring System. In *Proceedings of the 2011 ACM SIGMOD international conference on Management of data*, pages 1275–1278. ACM, 2011.

[63] I. Flaounas, N. Fyson, and N. Cristianini. Predicting relations in news-media content among EU countries. In *2nd International Workshop on Cognitive Information Processing*, pages 269–274. IEEE Press, Elba, Italy, 2010.

[64] I. Flaounas, M. Turchi, O. Ali, N. Fyson, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. The structure of EU mediasphere. *PLoS ONE*, page e14243, 2010.

[65] I. Flaounas, M. Turchi, and N. Cristianini. Detecting macro-patterns in the European mediasphere. In *IEEE/WIC/ACM Joint International Conference on Web Intelligence and Intelligent Agent Technology*, pages 527–530. Milano, Italy, 2009.

[66] I. Flaounas, M. Turchi, T. De Bie, and N. Cristianini. Inference and validation of networks. In *Machine Learning and Knowledge Discovery in Databases, European Conference (ECML PKDD)*, volume 5781 of *Lecture Notes in Computer Science*, pages 344–358. Springer, Bled, Slovenia, 2009.

[67] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

[68] B. Fortuna, C. Galleguillos, and N. Cristianini. *Detecting the bias in media with statistical learning methods*. Chapman & Hall/CRC press, 2009.

[69] R. Fowler. *Language in the News: Discourse and Ideology in the Press*. Psychology Press, 1991.

[70] B. Franklin. *Newszak and News Media*. Arnold, 1997.

[71] B. Franklin, M. Hamer, M. Hanna, M. Kinsey, and J. Richardson. *Key concepts in journalism studies*. Sage Publications Ltd, 2005.

[72] J. A. Fusaro and W. M. Conover. Readability of two tabloid and two non-tabloid newspapers. *Journalism Quarterly*, 60:141–144, 1983.

[73] C. G. What do you do with a million books? *D-Lib Magazine*, 12, 2006.

[74] J. Galtung and M. Ruge. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of International Peace Research*, 1:64–91, 1965.

[75] H. J. Gans. *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and Time*. Northwestern University Press, 25th anniversary edition edition, 2004.

[76] A. Ghose and S. Yang. An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10):1605–1622, 2009.

[77] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *Int. Conf. on Weblogs and Social Media (ICWSM 2007)*, 2007.

[78] M. González and A.-L. Barabási. Complex networks: From data to models. *Nature Physics*, 3:224–225, 2007.

[79] M. González, C. Hidalgo, and A. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[80] D. Gusfield. *Algorithms on strings, trees, and sequences*. Cambridge University Press, 1997.

[81] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[82] T. Harcup. *Journalism: Principles and Practice*. Sage, 2004.

[83] T. Harcup and D. O'Neill. What is news? Galtung and Ruge revisited. *Journalism Studies*, 2:261–280, 2001.

[84] I. Hargreaves. *Journalism: Truth or Dare?* Oxford University Press, 2003.

[85] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. *Annual Meeting – Association for Computational Linguistics*, 35:174–181, 1997.

[86] V. Hatzivassiloglou and J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the International Conference on Computational Linguistics*, pages 299–305, 2000.

[87] E. Hensinger, I. Flaounas, and N. Cristianini. Learning the preferences of news readers with SVM and Lasso ranking. In *6th IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 179–186. Larnaca, Cyprus, 2010.

[88] E. Hensinger, I. Flaounas, and N. Cristianini. Learning reader's preferences with ranking SVM. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 322–331. Springer, 2011.

[89] E. Herman and N. Chomsky. *Manufacturing consent : the political economy of the mass media*. Pantheon Books, New York, 1988.

[90] Hetherington. *News, Newspapers and Television*. Macmillan, London, 1985.

[91] A. Hirsh and H. Fraser. Protein dispensability and rate of evolution. *Nature*, 411:1046–1049, 2001.

[92] T. Hristo, V. Zavarella, J. Linge, M. Kabadjov, J. Piskorski, M. Atkinson, and R. Steinberger. Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *linguaMTICA Journal*, 2:55–66, 2009.

[93] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[94] K. Janes and M. Yaffe. Data-driven modelling of signal-transduction networks. *Nature Reviews Molecular Cell Biology*, 7:820–828, 2006.

[95] H. Jeong, S. Mason, A. Barabási, and Z. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[96] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *10th European Conference on Machine Learning (ECML)*, pages 137–142. Springer Verlag, 1998.

[97] T. Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer, 2002.

[98] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142. ACM, 2002.

[99] T. Joachims. A support vector method for multivariate performance measures. In *International Conference on Machine Learning (ICML)*, 2005.

[100] J. L. Johns and T. E. Wheat. Newspaper readability. *Reading World*, 18:141–147, 1987.

[101] I. Jordan, I. Rogozin, Y. Wolf, and E. Koonin. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Research*, 12(6):962–968, 2002.

[102] U. Kang, C. Tsourakakis, and C. Faloutsos. PEGASUS: A peta-scale graph mining system implementation and observations. In *International Conference on Data Mining*, pages 229–238. IEEE, 2009.

[103] E. Key, L. Huddy, M. Lebo, and S. Skiena. Large scale online text analysis using Lydia. In *American Political Science Association, Annual Meeting*, 2010.

[104] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[105] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86, 2005.

[106] P. Koehn, F. Och, and D. Marcu. Statistical phrase-based translation. In *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Association for Computational Linguistics*, pages 48–54, 2003.

[107] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[108] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8:159–178, 2005.

[109] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[110] G. Lanckriet, T. De Bie, N. Cristianini, M. Jordan, and W. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 2004.

[111] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Computational Social Science. *Science*, 323(5915):721–723, 2009.

[112] H. Lee, A. Smeaton, N. O'Connor, and B. Smyth. User evaluation of Físchlár-News: an automatic broadcast news delivery system. *ACM Transactions on Information Systems (TOIS)*, 24(2):145–189, 2006.

[113] J. Leskovec. *Dynamics of large networks*. Thesis, Carnegie Mellon University Pittsburgh, PA, USA, 2008.

[114] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[115] J. Leskovec and E. Horvitz. Planetary-scale views on an instant-messaging network. *Microsoft Research Technical Report, MSR-TR-2006-186*, 2007.

[116] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[117] J. Lewis. *Constructing Public Opinion*. Columbia Univerity Press, New York, 2001.

[118] J. Lewis, A. Williams, and B. Franklin. Four rumours and an explanation. *Journalism Practice*, 2(1):27–45, 2008.

[119] B. Liu. *Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2007.

[120] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop? In *AAAI Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.

[121] L. Lloyd, D. Kechagias, and S. Skiena. Lydia: A system for large-scale news analysis. *String Processing and Information Retrieval (SPIRE 2005)*, pages 161–166, 2005.

[122] L. Lloyd, A. Mehler, and S. Skiena. Identifying co-referential names across large corpora. In *Proc. Combinatorial Pattern Matching (CPM 2006)*, 2006.

[123] U. Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*, 2005.

[124] A. Ma'ayan. Insights into the organization of biochemical regulatory networks using graph theory analyses. *Journal of Biol. Chem.*, 284: 5451–5455, 2009.

[125] L. Manevitz and M. Yousef. One-class SVMs for document classification. *Journal of Machine Learning Research*, 2:139–154, 2001.

[126] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge Mass, 1999.

[127] M. Maragoudakis and D. Serpanos. Towards stock market data mining using enriched random forests from textual resources and technical indicators. In *Artificial Intelligence Applications and Innovations*, volume 339 of *IFIP*, pages 278–286. Springer, 2010.

[128] D. Matheson. Weblogs and the epistemology of the news: some trends in online journalism. *New Media & Society*, 6(4):443, 2004.

[129] M. McCombs and D. Shaw. The agenda-setting function of the mass media. *Public Opinion Quarterly*, 36:176–187, 1972.

[130] M. McCombs, D. Shaw, and D. Weaver. *Communication and Democracy: Exploring the Intellectual Frontiers of Agenda-setting Theory*. Mahwah, NJ:Lawrence Erlbaum Associates, 1997.

[131] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[132] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and

summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the second international conference on Human Language Technology Research*, page 285. Morgan Kaufmann Publishers Inc., San Diego, USA, 2002.

[133] S. McLachlan and P. Golding. *Tabloidization in the British Press: A Quantitative Investigation into Changes in British Newspapers, 1952–1997*. Rowman & Littlefield Pub Inc, 2000. 76-90 pp.

[134] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):765–772, 2006.

[135] O. Meyers, E. Zandberg, and M. Neiger. Prime time commemoration: An analysis of television broadcasts on Israel's memorial day for the holocaust and the heroism. *Journal of Communications*, 59:456–480, 2009.

[136] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[137] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *Arxiv cond-mat/0312028*, 2003.

[138] R. Milo, S. Shen-Orr, , S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[139] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[140] T. Mitchell. Mining our reality. *Science*, 326(5960):1644–1645, 2009.

[141] J. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[142] M. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103:8577–8582, 2006.

[143] J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[144] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[145] L. Paris and G. Bazzoni. The protein interaction network of the epithelial junctional complex: A system-level analysis. *Mol. Biol. Cell*, 19:5409–5421, 2008.

[146] C. Paterson. News agency dominance in international news on the Internet. *Papers in International and Global Communication*, 1:1752–1793, 2006.

[147] M. Pelillo. Replicator equations, maximal cliques, and graph isomorphism. *Neural Computation*, 11(8):1933–1955, 1999.

[148] M. Platakis, D. Kotsakos, and D. Gunopulos. Searching for events in the blogosphere. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1225–1226. ACM, New York, NY, USA, 2009.

[149] J. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, 1999.

[150] M. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[151] T. Potthast. Paradigm shifts versus fashion shifts? *EMBO reports*, 10: S42–S45, 2009.

[152] F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52:199–215, 2003.

[153] D. Radev, S. Blair-Goldensohn, Z. Zhang, and R. Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Proceedings of the first international conference on Human language technology research*, pages 1–4. Association for Computational Linguistics, 2001.

[154] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, S. Strogatz, and O. Sporns. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12): e14248, 2010.

[155] J. Rennie and A. Mccallum. Efficient web spidering with reinforcement learning. In *International Conference on Machine Learning*, 1999.

[156] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.

[157] E. Sandhaus. The New York Times annotated corpus. In *Linguistic Data Consortium*. Philadelphia, 2008.

[158] J. Savoy and L. Dolamic. How effective is Google's translation service in search? *Communications of the ACM*, 52:139–143, 2009.

[159] B. Scholkopf and A.Smola. *Learning with Kernels*. MIT Press, Cambridge Mass., 2002.

[160] B. Scholkopf, J. Platt, J.Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Technical report, Microsoft Research, MSR-TR-99-87*, 1999.

[161] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[162] J. G. Shanahan and N. Roma. Boosting support vector machines for text classification through parameter-free threshold relaxation. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 247–254. ACM, New York, NY, USA, 2003.

[163] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13:2498–2504, 2003.

[164] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[165] T. Snowsill, I. Flaounas, T. De Bie, and N. Cristianini. Detecting events in a million New York Times articles. In *Machine Learning and Knowledge Discovery in Databases, European Conference (ECML PKDD)*, pages 615–618. Barcelona, Spain, 2010.

[166] T. Snowsill, N. Fyson, T. De Bie, and N. Cristianini. Refining causality: who copied from whom? In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 466–474, 2011.

[167] T. Snowsill, F. Nicart, M. Stefani, T. De Bie, and N. Cristianini. Finding surprising patterns in textual data streams. In *Cognitive Information Processing, 2nd International Workshop on*, pages 405–410, 2010.

[168] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7: 1531–1565, 2006.

[169] R. Steinberger, B. Pouliquen, and E. V. der Goot. An introduction to the Europe Media Monitor family of applications. In *Information Access in a Multilingual World-Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009)*, pages 1–8, 2009.

[170] R. Steinberger, B. Pouliquen, et al. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Arxiv preprint cs/0609058*, 2006.

[171] G. H. Stempel. Readability of six kinds of content in newspapers. *Newspaper Research Journal*, 3:32–37, 1981.

[172] R. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[173] R. Swan and J. Allan. Extracting significant time varying features from text. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 38–45. ACM, 1999.

[174] G. Szabó and B. A. Huberman. Predicting the popularity of online content. *CoRR*, abs/0811.0405, 2008.

[175] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010.

[176] J. Szymański and W. Duch. Representation of hypertext documents based on terms, links and text compressibility. In *Neural Information Processing. Theory and Algorithms*, volume 6443 of *Lecture Notes in Computer Science*, pages 282–289. Springer Berlin / Heidelberg, 2010.

[177] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: A large margin approach. In *22nd International Conference on Machine Learning*, 2005.

[178] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

[179] K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259, 2003.

[180] K. Toutanova and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 63–70, 2000.

[181] G. Trunk. A problem of dimensionality: A simple example. *Trans. on Pattern Analysis and Machine Intelligence*, 1:306–307, 1979.

[182] M. Turchi, T. De Bie, and N. Cristianini. An intelligent web agent that autonomously learns how to translate. *Web Intelligence and Agent Systems: An International Journal (WIAS)*, 2011.

[183] M. Turchi, I. Flaounas, O. Ali, T. De Bie, T. Snowsill, and N. Cristianini. Found in translation. In *Machine Learning and Knowledge Discovery in Databases, European Conference (ECML PKDD)*, volume 5782 of *Lecture Notes in Computer Science*, pages 746–749. Springer, Bled, Slovenia, 2009.

[184] P. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics*, pages 417–424, 2002.

[185] R. Uribe and B. Gunter. Research note: The tabloidization of British tabloids. *European Journal of Communication*, 19(3):387, 2004.

[186] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[187] A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.

[188] M. Wall. 'Blogs of war': Weblogs as news. *Journalism*, 6(2):153–172, 2005.

[189] D. Watts. A twenty-first century science. *Nature*, 445(7127):489, 2007.

[190] F. Wu and B. A. Huberman. Popularity, novelty and attention. In *Proceedings 9th ACM Conference on Electronic Commerce (EC-2008)*, pages 240–245, 2008.

[191] X. Wu, R. Srihari, and Z. Zheng. Document representation for one-class SVM. *Machine Learning: ECML 2004*, pages 489–500, 2004.

[192] Y. Yang and X. Liu. A re-examination of text categorization methods. In *In Proceedings of the 22nd annual international ACM SIGIR*

conference on research and development in information retrieval, pages 42–49. ACM Press, 1999.

[193] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 28–36. ACM, 1998.

[194] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Conference on Emprical Methods in Natural Language Processing*, pages 129–136, 2003.

[195] D. Zhang and S. Simoff. Informing the curious negotiator: Automatic news extraction from the internet. In *Data Mining*, pages 176–191. Springer, 2006.

[196] W. Zhang and S. Skiena. Trading strategies to exploit blog and news sentiment. In *Fourth Int. Conf. on Weblogs and Social Media (ICWSM)*, 2010.