SENTIMENT-FOCUSED WEB CRAWLING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

AVNI GÜRAL VURAL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

SEPTEMBER 2013

Approval of the thesis:

SENTIMENT-FOCUSED WEB CRAWLING

submitted by **AVNI GÜRAL VURAL** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Assoc. Prof. Dr. Pınar Karagöz _____
Supervisor, **Computer Engineering Department, METU**

Dr. Berkant Barla Cambazoğlu _____
Co-supervisor, **Yahoo Labs, Barcelona**

Examining Committee Members:

Prof. Dr. Özgür Ulusoy
Computer Engineering Department, Bilkent University _____

Assoc. Prof. Dr. Pınar Karagöz
Computer Engineering Department, METU _____

Prof. Dr. İ. Hakkı Toroslu
Computer Engineering Department, METU _____

Prof. Dr. Nihan Kesim Çiçekli
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Hakan Ferhatosmanoğlu
Computer Engineering Department, Bilkent University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: AVNI GÜRAL VURAL

Signature :

ABSTRACT

SENTIMENT-FOCUSED WEB CRAWLING

Vural, Avni Güral

Ph.D., Department of Computer Engineering

Supervisor : Assoc. Prof. Dr. Pınar Karagöz

Co-Supervisor : Dr. Berkant Barla Cambazoğlu

September 2013, 94 pages

The advent of Web 2.0 has led to an increase in the amount of sentimental content available in the Web. Such content is often found in social media web sites in the form of product reviews, user comments, testimonials, messages in discussion forums, status updates, and personal blogs as well as in other forms, including opinions in personal pages, news articles, and product descriptions. The analysis of sentimental content has a number of important applications, most important being web search, contextual advertisement, and recommendation. The timely discovery of sentimental content is important as most sentiments quickly lose their value if they are not immediately discovered. So far, all focused crawlers work in a topic-specific manner and fall short when sentimental pages are focused to be discovered. In addition, up to date, most of the research carried on sentiment analysis was focused on English language.

In this thesis, we present a new perspective for focused web crawling. First, we propose a sentiment-focused web crawling framework to facilitate the quick discovery of sentimental content and evaluate it via simulations over the publicly available ClueWeb09-B web page collection. Second, we propose a framework for unsupervised sentiment analysis in Turkish and perform experiments with data from popular Turkish social media sites. Finally, we consolidate our frameworks and present a customized version of sentiment-focused web crawling framework for Turkish.

Keywords: focused web crawling, sentiment analysis, sentimentality, polarity, Turkish, ClueWeb09

ÖZ

DÜŞÜNCE ODAKLI WEB TARAYICILIK

Vural, Avni Güral
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Pınar Karagöz
Ortak Tez Yöneticisi : Dr. Berkant Barla Cambazoğlu

Eylül 2013 , 94 sayfa

Web 2.0'ın gelişi, Web üzerinde yer alan düşünce ve görüş içeriğinin artmasını sağlamıştır. Düşünce ve görüş içeren içerikler sıkça sosyal medya sitelerinde ürün yorumu, kullanıcı yorumu, tartışma forum mesajı, durum güncellemesi, kişisel blog sayfası, haber sayfası, ürün açıklaması şeklinde bulunmaktadır. Düşünce içeriklerinin analizinin web arama, içeriksel reklam, tavsiye gibi bir çok önemli uygulaması mevcuttur. Düşünce içeriklerinin çoğu hemen keşfedilmezse değerlerini kaybetmektedir, bu nedenle bu içeriklerin zamanında keşfi çok önemlidir. Şu ana kadar bütün odaklı web tarayıcılar konu odaklı çalışmaktadır ve bu yaklaşım düşünce içeren sayfaların keşfedilmesinde yetersiz kalmaktadır. Ayrıca, bu zamana kadar düşünce analizi ile ilgili yürütülen araştırmalar ağırlıklı olarak İngilizce diline odaklanmıştır.

Bu tezde odaklı web tarayıcılığa yeni bir perspektif getirilmektedir. İlk önce, düşünce içeren sayfaların daha hızlı keşfini sağlayan bir düşünce odaklı web tarayıcı çatısı önerilecek ve ClueWeb09-B web sayfası koleksiyonu üzerinde simülasyonlar ile değerlendirmeler yapılacaktır. İkinci olarak, Türkçe düşünce analizi çatısı önerilecek ve popüler Türkçe sosyal medya site verileri üzerinde deneyler gerçekleştirilecektir. En son olarak, önerilen bu iki çatı birleştirilerek Türkçe için düşünce odaklı web tarayıcı çatısı sunulacaktır.

Anahtar Kelimeler: Odaklı web tarayıcılık, düşünce analizi, hissiyat, tutum, Türkçe, ClueWeb09

Dedicated to my dear wife Nur and our lovely children Doğa and Ada.

ACKNOWLEDGMENTS

This was a long journey starting from 2001 although I had a break from 2005 to 2011. During this journey, I had many challenges, sleepless nights, unstable feelings, deadlines, at least one failure (in 2005), successes... It was all worth it.

First of all, I would like to express my deepest gratitude to my thesis supervisors Dr. Pınar Karagöz and Dr. Berkant Barla Cambazoğlu. Their guidance, broad knowledge and critical thinking have been of great value for me throughout this research. I really feel lucky that I was able to study with them.

I would like devote another paragraph for Dr. Berkant Barla Cambazoğlu, the idea owner of this thesis topic. He is an exceptional scientist and I really learnt a lot from him. Words cannot express my gratitude to him.

I would like to thank Prof. Dr. Özgür Ulusoy and Prof. Dr. Hakkı Toroslu for their valuable suggestions and comments throughout the steering meetings of this study. I would also like to thank to Prof. Dr. Nihan Kesim Çiçekli and Assoc. Prof. Dr. Hakan Ferhatosmanoğlu for kindly accepting being in the examining committee.

I also acknowledge and thank to Ertan Özgür, Selma Süloğlu and Ebru Vural for their participation in user-study; Gürkan Vural, Turgay Çelik, Erkin Çilden, Deniz Oğuz and Alev Mutlu for their invaluable discussions; Nazlı İkizler Cinbiş for proof reading; Umut Eroğul and Mesut Kaya for sharing their dataset used in [38] and [56] respectively; and MilSOFT for its support to academic studies.

I would like to acknowledge that this research has been partially supported by TÜBİTAK (Scientific and Technical Research Council of Turkey) with grant number 112E002.

Finally, beyond and above all, I would like to thank my family. None of this would be possible, without their unconditional love and endless support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ASCII	American standard code for information interchange
AP	Average precision
API	Application programming interface
DCG	Discounted cumulative gain
DNS	Domain name system
DOM	Document object model
FIFO	First in first out
GT	Ground truth
HTML	Hyper text markup language
IMDB	Internet movie database
KNN	K-nearest neighbors
NLP	Natural language processing
RAM	Random access memory
SVM	Support vector machine
TREC	Text retrieval conference
URL	Uniform resource locator
WWW	World wide web

CHAPTER 1

INTRODUCTION

"Many of life's failures are people who did not realize how close they were to success when they gave up."

– Thomas Edison

1.1 Motivation

Other people's opinions have always been invaluable information for us, especially in decision making process. People seek for advices for recommendation of a doctor or a mobile phone or even a politician for voting. Organizations conduct opinion polls and surveys in order to gather public opinions about their products and services. Nowadays the main source for opinions and sentiments is the World Wide Web (WWW) as there are numerous pages for blogs, news articles, product reviews, user comments, testimonials, professional critics, discussion forums etc. It is a fact that Web hosts a tremendous amount of data in various forms. Figure-1.1 summarizes what happens on Web just in one minute. 571 new web sites creation, over 100.000 tweets sent on *Twitter*[1], 684,478 pieces of content share on *Facebook*[2], over 2 million search queries on *Google*[3], and 3,125 photo uploads on *Flickr*[4] are some of the striking statistics for internet activities per minute. *WordPress*[5], a well-known weblog hosting provider, hosts more than 67 million individual blogs with the service as of August 2013 and reports that [4]:

- As of 2011, over 100,000 new WordPresses are created every day.
- WordPress.com users produce about 1.5 million new posts and 2 million new comments on an average day (which was only 500,000 new posts and 400,000 new comments per day as of August 2012).

[1] Twitter, `http://www.twitter.com`
[2] Facebook", `http://www.facebook.com`
[3] Google, `http://www.google.com`
[4] Flickr, `http://www.flickr.com/`
[5] WordPress, `http://www.wordpress.com`

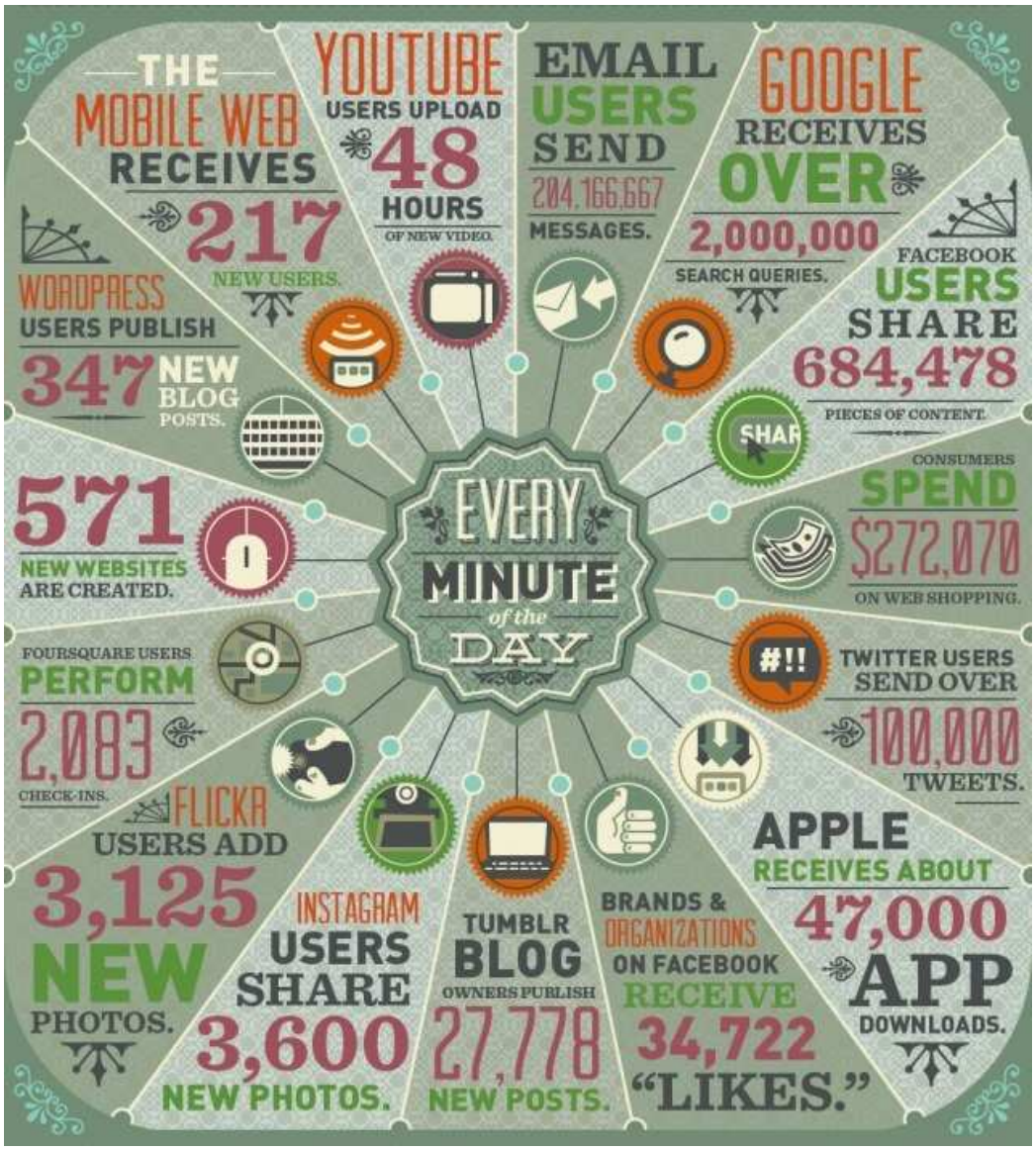


Figure 1.1: What happens in an Internet minute.[6]

- Over 363 million people view more than 10.7 billion pages each month (which was 2.5 billion pages per month as of August 2012).
- WordPress blogs are written in over 120 languages. The majority of blogs with rate 66%, are in English.

Timely discovery of the sentimental or opinionated web content has a number of advantages. The most important of all, it has a high potential for monetization. Understanding of the sentiments of human masses towards different entities and products enables better services for contextual advertisement, recommendation systems and analysis of

[6] `http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/?dkw=socf3`.

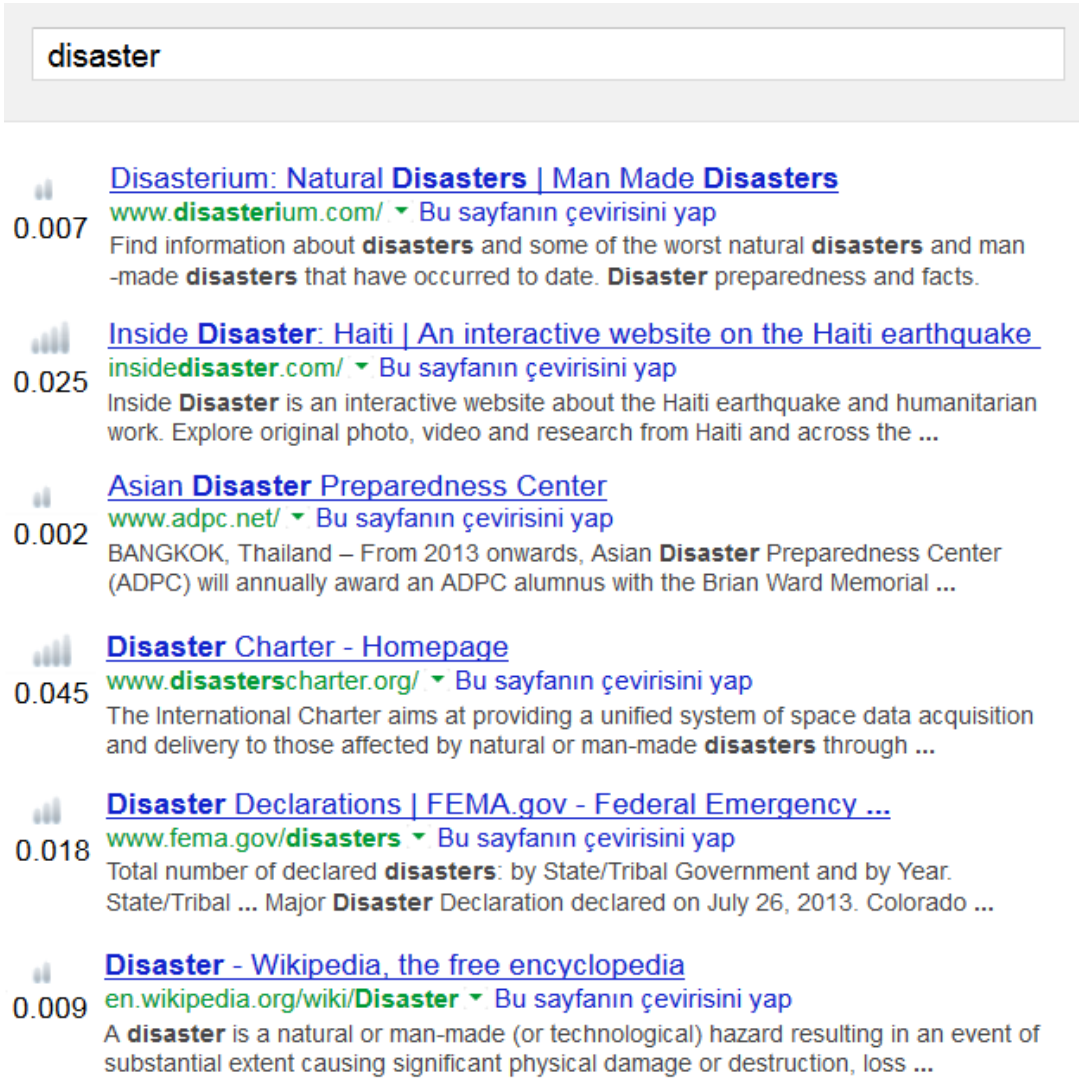


Figure 1.2: Mock-up search engine listing results with sentimentality scores.

market trends. Another advantage is enabling for a sentimental search engine by locating the required web content for indexing. Demartini and Siersdorfer [35] report that no extreme sentiments are shown in the top results of three popular search engines. Figure 1.2 shows a mock-up user interface for a search engine listing the search results with a numerical and/or graphical indicator for how sentimental the corresponding content is.

In the recent years, there has been a growing interest on analysis of opinionated context in terms of sentiment analysis and opinion extraction [40], mostly for English language [72]. However, the automatic discovery of the sentimental or opinionated web content is mostly ignored in the literature. In addition, because of the proliferation of diverse sites, finding and monitoring sites with sentimental or opinionated content on the Web and distilling the contained information remains a challenging task [63].

Crawlers, a search engine's crucial component for collecting web resources [15], have a key role in content discovery on Web. Given the magnitude of the data on the Web, crawlers need to prioritize and crawl Web pages by focusing only on the "valuable" regions of Web. Such *focused crawlers* work in a topic-specific manner, by predicting the probability that a link to a particular page is relevant to a given topic before actually downloading the page [27]. This strategy falls short when sentimental pages (i.e. web pages containing sentiment/opinion, such as blogs, news articles etc.) are the primary target of the discovery process, which requires a relevance or ranking metric for sentimentality.

This thesis focuses on the *sentiment-focused web crawling* problem and proposes a crawling framework for faster discovery and retrieval of sentimental sources on the Web. The proposed framework is further extended for Turkish language, for which there exists limited number of academic works on sentiment analysis.

1.2 Contribution

This thesis makes the following contributions.

- We make a sentiment-focused web crawler design, using the state-of-the-art page processing and sentiment analysis tools. We propose techniques for predicting the sentimentality of unseen web pages and their prioritization by the crawler.
- We empirically evaluate our techniques against traditional crawling strategies via simulations over the publicly available ClueWeb09-B web page collection, using different seed page selection strategies.
- We conduct a user study that facilitates obtaining a meaningful ground-truth for the sentimentality of the web pages in the ClueWeb09-B collection.
- We extend our sentiment-focused web crawler design to focus on polarity to discover pages including more positive content as early as possible.
- We propose a framework for unsupervised sentiment analysis in Turkish by customizing a state-of-the-art sentiment analysis library.
- We further extend our sentiment-focused web crawler design to crawl for Turkish web pages, utilizing our sentiment analysis framework proposed.

The remainder of this thesis is organized as follows.

In Chapter 2 we provide a brief survey of related work in the context of web crawling, focused web crawling, sentiment analysis, and focused web crawling that involve sentimentality.

In Chapter 3, we propose an elegant design for sentiment-focused web crawling, using the state-of-the-art page processing and sentiment analysis tools [87]. Within this framework, we also propose different techniques for predicting the sentimentality of "unseen" web pages and their prioritization by the crawler. We conduct a user study to form ground-truth for sentimentality scores of web pages in our web page collection. We report experiment results over a large collection of web pages, with use of different seed selection techniques. We further extend our framework for polarity focused web crawling.

In Chapter 4, we propose a framework for unsupervised sentiment analysis in Turkish by customizing a state-of-the-art sentiment analysis library, SentiStrength. The first version of this framework is proposed by us in [88]. We provide experiment results on classifying the polarity of texts including movie reviews, hotel reviews and political news, obtained from popular Turkish social media sites.

In Chapter 5, we modify our sentiment-focused web crawler proposed in Chapter 3, to crawl for Turkish web pages, utilizing the sentiment analysis framework proposed in Chapter 4. We create our own web collection, using more than 400 hundred seed pages in Turkish. We conduct a user study to form ground-truth for sentimentality scores of web pages in our web page collection. We report experiment results both for sentimentality and polarity focus, applying different seed selection techniques.

Finally, we conclude and point to some future work directions in Chapter 6.

CHAPTER 2

RELATED WORK

"The significant problems we face cannot be solved at the same level of thinking we were at when we created them."

– Albert Einstein

This chapter provides a brief survey of related work in the context of web crawling (Section 2.1), focused web crawling (Section 2.2), sentiment analysis (Section 2.3), and focused web crawling with sentiment exploitation (Section 2.4).

2.1 Web Crawling

A *web crawler*, also known as harvester, spider, or robot, is an application that browses the World Wide Web and automatically downloads web pages. The most widespread use of crawlers is in support of search engines by providing data for building indexes. Crawlers can also be used for gathering specific information (for instance e-mail addresses) from web pages or automating site maintenance (for instance by checking links).

Figure-2.1 shows how a basic crawler works. The main part of such a basic crawler is the *download queue*, (or sometimes referred as *frontier*). Download queue maintains a list of unvisited Uniform Resource Locators (URL) and initially includes seed URLs which may be provided by the user or another application. During the execution of the crawler, first a URL from the download queue is picked in, after that the page corresponding to the URL is fetched and stored in Content Database, and finally new links extracted from the fetched page are added to the queue. The process repeats until either a certain number of pages have been crawled, or the download queue becomes empty.

When the download queue is implemented as a FIFO queue, the pages are fetched in a breadth-first manner. This crawling strategy is explored in the first comprehensive full-text search engine, WebCrawler [75], and as well as in a more recent research [31].

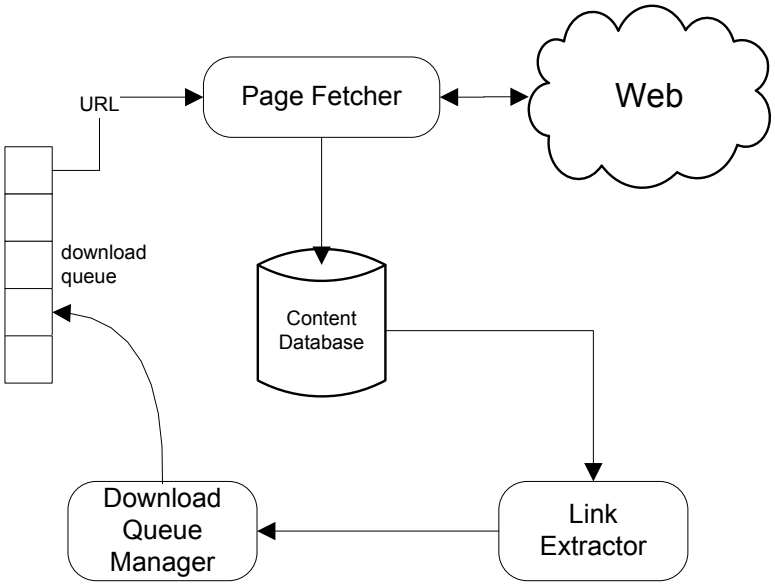


Figure 2.1: Basic web crawling architecture.

Web crawling has many challenges. As stated by Najork [70], all challenges are ultimately related to scale in deed. A typical crawler has to fetch *thousands of pages in a second* in order to maintain a fresh corpus of search engine with a couple of billion pages.

- **Content Selection**: Crawlers should bypass irrelevant, low-quality, malicious content.
- **Scale**: Web hosts an enormous amount of data but the crawlers seek broad coverage and freshness.
- **Speed**: The basic and fast way of keeping track of visited URLs is storing in the memory. However, it is not possible to store hundreds of millions of URLs in the memory. That is why such lists are kept on disk, but accessing and re-ordering the list becomes slower which degrades the crawling performance.
- **Infrastructure Cost**: According to a blog post on the Official Google Blog [9], the Google index now contains 1 trillion unique URL's. This graph of one trillion URLs is similar to a map made up of one trillion intersections, which is continuously reprocessed several times per day. Storing, indexing and fresh-keeping such a huge data requires massive investments on hardware infrastructure.
- **Ethics / Politeness**: Crawlers should prevent server overload and comply with Robot Exclusion Protocol (which is a de facto standard for definition of access rights on a website for crawling) [59].

2.2 Focused Web Crawling

For sure, the most important property for a crawler is the speed. The web, hosting an enormous amount of data in various forms, is estimated to have more than 10 billion pages and continues to grow rapidly [2]. Crawlers should discover and use not only the new pages but also the recent versions of the pages. A general purpose crawler is said to be neither necessary nor sufficient to crawl and index such a giant amount of web pages. Therefore, a specialized version of crawling, called *focused web crawling*, is proposed aiming to collect on-topic data which actually reduces the search space.

2.2.1 Early Algorithms

All the early algorithms for focused crawling rely on statistical techniques for estimating page relevance without using a classifier.

The first related work on focused crawling was done by De Bra et al. [22], called *fish search*. In this research, for a client-based search engine, a simulated *group of fish* crawls on the web and assigns page relevance metric to pages using a binary classification with an input of simple keyword or regular expression. The lifespan of a fish depends on visited page relevance, where fish dies when a specified amount of irrelevant pages are traversed. In addition, a fish produces offspring when a new page is traversed. Fish search works as depth-first search, since new URLs are added to the beginning of the download queue.

Later, an extended version of fish search, so called *shark search* was proposed by Hersovici et al. [48]. In contrast to fish search, shark search estimates page relevance with a continuous function resulting a real number between 0 and 1 for the a similarity between document and query. Then, in the download queue, URLs are ordered by linear combination of source page relevance, anchor text and neighbourhood of the link on the source page.

2.2.2 Focused Crawling with Classifier

After the preliminary works, Chakrabarti et al. [27] first introduced a focused crawler based on a classifier. According to Chakrabarti et al. in [27], a focused crawler seeks pages on a specific set of topics that represent a relatively narrow segment of the Web, guided by a text classifier, using an existing document taxonomy (e.g. pages in Yahoo tree) for training phase. Thus, a focused crawler implements a *best-first* search strategy, rather than the breadth-first search applied by general purpose crawlers. Figure-2.2 illustrates the basic architecture for a focused crawler.

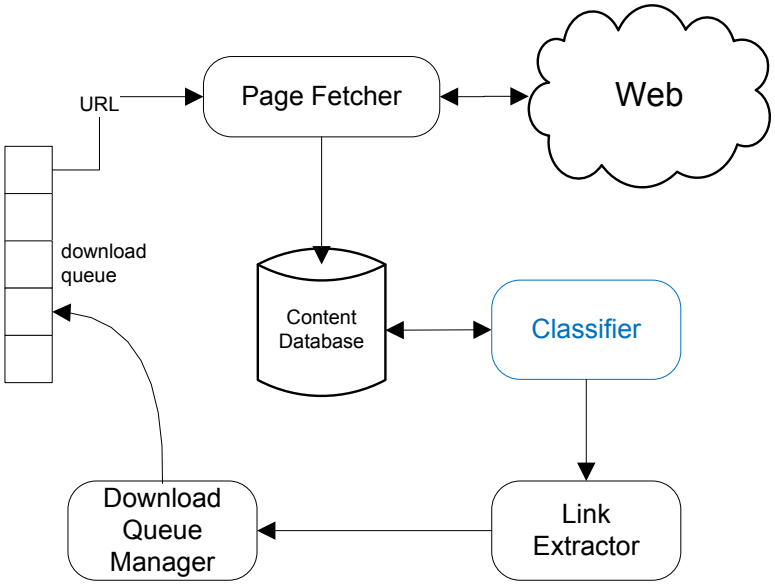


Figure 2.2: Basic focused web crawling architecture.

As Chakrabarti et al. [27] defines focused crawler with three main components: a *classifier*, a *distiller* and a *crawler*. The classifier and distiller govern the crawler as the classifier determines the relevance/importance of the crawled pages to decide on the link expansion and the distiller computes a measure of the centrality of crawled pages to determine their download priorities. Chakrabarti et al. present two different rules for link expansion, which are the "hard focus" rule, allowing expansion of links only if the class to which the source page belongs with the highest probability is in the interesting subset, and the "soft focus rule", using the sum of probabilities that the page belongs to one of the relevant classes to decide the visit priority of child pages. In a more recent study of Chakrabarti et al. [26], page relevance and URL visit priorities are decided by separate models. In both studies, Chakrabarti et al. empirically show that a focused crawler successfully stays within scope, and explores a steadily growing population of topical pages over time.

In a study of Menczer et al. [68], best-first search, PageRank [23], and InfoSpider [67] focused web crawlers are evaluated based on ability to remain on-topic during the crawling session. PageRank based crawler maintains a download queue which priorities the links with respect to their PageRank score whereas InfoSpider uses a set of agents, with a back-propagation neural network in deciding link to follow. Menczer et al. [68] concluded that best-first outperformed the other two where PageRank finished last. There are several possible explanations for this result. First of all, unsurprisingly, the topic similarity metric defined for best-first is the key for its success. On the contrary, PageRank method favours authoritative pages in a general context, that is why it is too general for topic-specific search.

It is necessary to derive a powerful model for the classifier since the performance of a focused web crawler depends on the model learnt. Most studies employ various machine learning techniques, such as hidden Markov models [18], conditional random fields [64], reinforcement learning [66], genetic algorithms [54], and support vector machines [32]. In order to improve the learned models, ontologies [37] and link semantics [91] are also exploited. Li et al. [62] proposes a focused crawler guided by anchor texts using a decision tree for prioritizing URLs.

The focused web crawling inherit the challenges of web crawling as well. In addition, focused web crawling has some weakness, such as tunnelling, which is the inability to learn that a path of off-topic pages can eventually lead to on-topic pages. To solve the problem of tunnelling, Diligenti et al. [36] use context graphs while Altingovde et al. [10] propose a rule-based crawler, which priorities the links using the rules derived from interclass linkage patterns.

The focused web crawling paradigm has been employed in a number of systems for gathering topic/domain specific Web pages, such as crawling computer science research papers for Cora search engine [77], biomedical applications [81], tourism and health pages in Indian language [76].

2.3 Sentiment Analysis

Opinions as the key influencers of our behaviours are important. As Liu stated in [63], our beliefs and perceptions of the reality, and the choices we make, are to a considerable degree conditioned on how others see and evaluate the world. This is the reason why people seek out opinions of others when making a decision. *Sentiment analysis* (a.k.a. *opinion mining*) is an area of study analysing people's opinions, appraisals, attitudes, feelings, and emotions toward entities, individuals, issues, events, topics, and their attributes. Sentiment analysis has been an active research area for quite some time. Feldman in [40] states that there are over 7,000 articles on the topic of sentiment analysis.

2.3.1 Formal Definition of Sentiment

Despite the slightly different meanings, people tend to use term sentiment interchangeably with terms of feeling, opinion, and emotion. Webster [1] defines sentiment as "*an attitude, thought, or judgement prompted by feeling*" while feeling is "*an emotional state or reaction*" and emotion is "*a state of feeling*".

As Liu defined in [63], "an *opinion* is simply a positive or negative *sentiment*, view, attitude, emotion, or appraisal about an entity or an aspect of the entity from an opinion

[1] Webster, `http://www.webster.com`

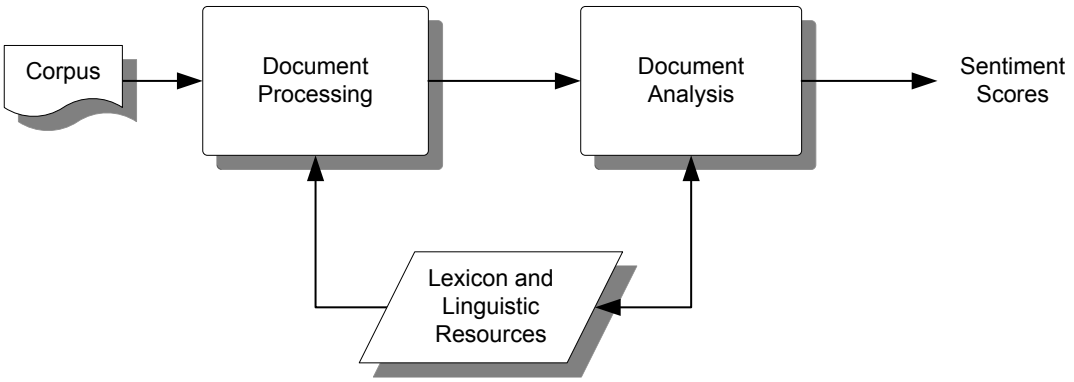


Figure 2.3: The pipeline of modules in a generic sentiment analysis framework.

holder". With the formal definition of Liu [63], an opinion is a quintuple, $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$, where e_i is the name of an entity, a_{ij} is an aspect of e_i, oo_{ijkl} is the orientation of the opinion about aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. Of course, some other attributes (such as gender, age, web site etc.) can also be added to the tuple. The opinion orientation oo_{ijkl} can be positive, negative, or neutral or be expressed with different strength/intensity levels. With this formulation, objective of sentiment analysis is discovering all opinion quintuples $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ in a collection of opinionated documents $\mathcal{D}$. This formal definition can also be further extended by adding more components to the tuple, such as gender, age, web-site, etc. but this makes the definition more problematic since all the components are essential for any analysis.

An example can be helpful to illustrate the formulation of an opinion. When we analyse the sentence "*The camera of my phone is great.*" written on a blog of *UserX* on *1 November 2011*, the target entity is "*phone*", the aspect is "*camera*", the opinion is "*positive*" due to sentimental term "*great*". Then the opinion quintuple will be as (*phone, camera_ quality, positive, UserX, 1-Nov-2011*).

2.3.2 Architecture for Sentiment Analysis

As Feldman stated in [40], apart from the techniques applied, all sentiment analysis systems implement a generic pipeline, which is illustrated in Figure 2.3. The input for a sentiment analysis framework is a corpus of documents, which may include a web page, a short post, or a text document in any format. The corpus is pre-processed by document processing module, which converts corpus to texts by using linguistic resources for stemming, tokenization etc. Then pre-processed texts are sent to the document analysis module which annotates them using the linguistic resources and sometimes with a dictionary of words, called *lexicon*, which includes annotations of

word's sentiment strength. The primary source of subjective content in a textual content can be adjectives [47, 51]; adverbs [21]; adjectives and verbs [58]; or exclusive use of verbs [80]. The output of the framework is the annotations which can be attached to whole document, to sentences, to entities as sentiment scores.

The document analysis module, which is the main module of the framework, can apply two main lines of techniques for annotation: machine-learned and lexicon-based. In machine-learned sentiment analysis techniques [73], a model is built by a training dataset using common machine learning algorithms such as SVM, Naive Bayes, Logistics Regression, or KNN. Each document instance (e.g., a product review) in the dataset is associated with a value indicating the strength and polarity of the sentiments expressed in the text. In practice, obtaining large-scale labeled datasets is difficult, and the existing datasets are often domain-specific. As a remedy, the lexicon-based sentiment analysis techniques [14, 84, 82] aim to create a vocabulary and a set of rules to quantify the amount of sentiments in a given piece of text, eliminating the dependency to labeled training data. The vocabulary and rules are created by manually, or expanded by utilizing resources like WordNet [69].

2.3.3 Tasks in Sentiment Analysis

Much of the early research on sentiment analysis focused on sentiment classification at the document and sentence levels in terms of subjectivity [72], polarity of subjectivity as positive or negative [86], and strength of subjectivity.

For sentiment classification, most of the works used product and movie reviews. Dave et al. [34] classified product reviews as positive or negative by using information retrieval techniques for feature extraction and scoring. They reported that best methods works as well as or better than traditional machine learning techniques, including SVM and Naive Bayes. Turney [86] used an unsupervised learning technique based on the estimated semantic orientation of extracted phrases to classify the movie reviews as "recommended" or "not recommended". A prediction accuracy of 65.8% is reported for a collection of 120 movie reviews. Pang et al. [73] compared the performance of different machine learning techniques on movie reviews taken from the IMDB [2] movie database. The SVM classifier is shown to yield better performance than the other classifiers. The used features included unigrams, bigrams, part of speech information, and the position of the terms in the text. Among these feature types, unigrams were found to yield better performance. In a recent work, Oghina et al. [71] tried to predict the movie ratings based on the feedback obtained from different social media channels like Twitter and YouTube. Kennedy and Inkpen [57] combined machine learning with a simple technique based on counting the positive/negative words in the movie reviews, showing further improvements over both techniques.

[2] IMDB, `http://www.imdb.com`

Table2.1: Sample entries for SentiStrength's sentiment word list

Word	Score	Word	Score
accomplish*	+2	blackmail*	−3
agreeab*	+1	brokenhearted*	−5
bliss	+5	crawl*	−1
glory	+2	scariest	−4
passion*	+3	sick*	−2
sentimental*	+3	ugl*	−3

The other tasks in sentiment analysis include, but not limited to, sentiment retrieval [44, 93, 29], sentiment extraction [90, 16], sentiment summarization [20, 61], aspect-based opinion summarization [85], comparative opinion mining [43], opinion spam detection [53, 52].

2.3.4 A Sentiment Analysis Tool - SentiStrength

SentiStrength[3] developed by Thelwall et. al ([84, 83]) is one of the powerful sentiment analysis tools for English in the literature. The tool assigns both positive and negative scores to words, having a range from +1 (not positive) to +5 (extremely positive) for positive sentiment strength scores and −1 (not negative) to −5 (extremely negative) for negative sentiment strength scores. A final judgement for the whole sentence is computed by taking the maximum and minimum scores among all individual positive and negative scores respectively.

The first version of SentiStrength has been applied to MySpace[4] comments. The model has been trained by a set of 2,600 human-classified MySpace comments, and evaluated on a further random sample of 1,041 MySpace comments. Thelwall et al. [84] reports that the tool is able to predict positive emotion with 60.6% accuracy and negative emotion with 72.8% accuracy, both based upon strength scales of 1-5. In 2012, an improved version of SentiStrength is announced[83] having successful performance on six diverse social web data sets (MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums).

The tool defines its corpus in following types of lists:

- A **sentiment word list** with human polarity and strength judgements (e.g. support 2), having more than 2500 entries, each associated with a sentiment score between −5 and +5. Some of the words include Kleene star stemming (e.g., "amaz*", which covers "amaze", "amazed", "amazement", "amazing", "amazingly" etc.). Sample entries for sentiment word list are given in Table 2.1.

[3] SentiStrength, `http://sentistrength.wlv.ac.uk`.
[4] MySpace, `http://www.myspace.com`

Table2.2: Sentiment scores generated by SentiStrength for sample English sentences

Sentence	Positive score	Negative score	Binary prediction
I can play chess	+1	−1	+1
I can play chess!!!	+2	−1	+1
I like to read science fictions	+2	−1	+1
I do not like to read science fictions	+1	−1	+1
I left early because the film was boring	+1	−2	−1
I hate you	+1	−4	−1
I really love you, but dislike your cold sister	+4	−3	+1

- A **booster word list** for strengthen or weaken the emotion of following sentiment words (e.g. extremely 2).
- An **idiom list** for identifying the sentiment of a few common phrases while overriding individual sentiment word strengths (e.g. how are you 2).
- A **negating word list** for inverting following emotion words (e.g. cannot).
- An **emoticon list** with polarities for identifying additional sentiment (e.g. :) 1).

Table 2.2 lists some sample English sentences and the sentiment scores produced by SentiStrength. "Binary prediction" column of Table 2.2 shows the polarity of subjectivity as positive or negative.

Further details about the tool are available in [84] and [83].

2.3.5 Sentiment Analysis in non-English Languages

Although most sentiment analysis techniques developed so far are for English, there are also studies for other languages in recent years. Atteveldt et al. [13] used machine learning techniques to automatically determine the polarity of political news stories in Dutch. They extracted lexical and syntactic features besides three different clusterings of similar words based on annotated material. Ghorbel and Jacot [45] devised a supervised learning strategy using linguistic features obtained through part-of-speech tagging and chunking as well as semantic orientation of words obtained from the SentiWordNet sentiment analysis tool [14] to classify the polarity of movie reviews in French. Since SentiWordNet is for English, the authors translated the French words to English before getting their semantic orientation. Sentimatrix developed by Gînscă et al. [46], performs a sentiment extraction and analysis with name entity recognition for multiple languages, currently for English and Romanian. Brooke et al. [24] adapted

an existing English sentiment analysis tool to Spanish and compared it with alternative approaches, including machine translation and machine learning. Freitas and Vieira [41] identified the polarity in Portuguese user generated movie and hotel reviews according to features described in domain ontologies. Zhang et al. [92] addressed the challenges that are unique to the Chinese language. They evaluated a rule-based polarity classification approach against different machine learning approaches. Hiroshi et al. [49] developed a sentiment analysis system using a transfer-based machine translation engine and applied it to Japanese. Abbasi et al. [6] studied sentiment analysis on English and Arabic content in Web forums. In a recent study, Abdul-Mageed et al. [7] built a subjectivity and sentiment analysis tool for modern standard Arabic. For multiple languages Bautin et al. [19] and for Chinese Wan [89] performed sentiment analysis by translating the texts into English and evaluating with an existing English sentiment analysis tool. Balahur and Turchi [17] performed multilingual sentiment analysis on French, German and Spanish using machine translation and concluded that for languages with high translation quality performance of sentiment analysis is similar to systems implemented for English.

In literature, the research on sentiment analysis in Turkish is limited. The first detailed analysis is presented in Erogul's master thesis [38], which rely on supervised machine learning for polarity classification. Later, Vural et al. [88] proposed a sentiment analysis framework for Turkish and apply it to the problem of classifying the polarity of movie reviews. Although the proposed framework is unsupervised, it is demonstrated to achieve a fairly good classification accuracy, approaching the performance of supervised polarity classification techniques. Recently, Kaya et al. [56] studied to clasify political news from columns in different Turkish news sites as supporting (positive) or criticizing (negative) using four supervised machine learning algorithms of Naïve Bayes, Maximum Entropy, SVM and character based N-Gram Language Model.

2.4 Focused Web Crawling with Sentiment Exploitation

In the literature, the automatic discovery of the sentimental web content is mostly ignored. In addition to the concurrent work of Fu et al. [42], which is the only prior work that exploits the sentiment information in focused web crawling, there are a number of initiatives for the sentimental or emotional search engines in the literature.

2.4.1 Graph-based Sentiment Crawler

Fu et al. [42] propose a focused crawler, called graph-based sentiment crawler, using a graph-based tunneling mechanism and a text classifier that incorporates topic and sentiment information for early discovery of sentimental content about a given topic. The high-level architecture of graph-based sentiment crawler adopts and extends the

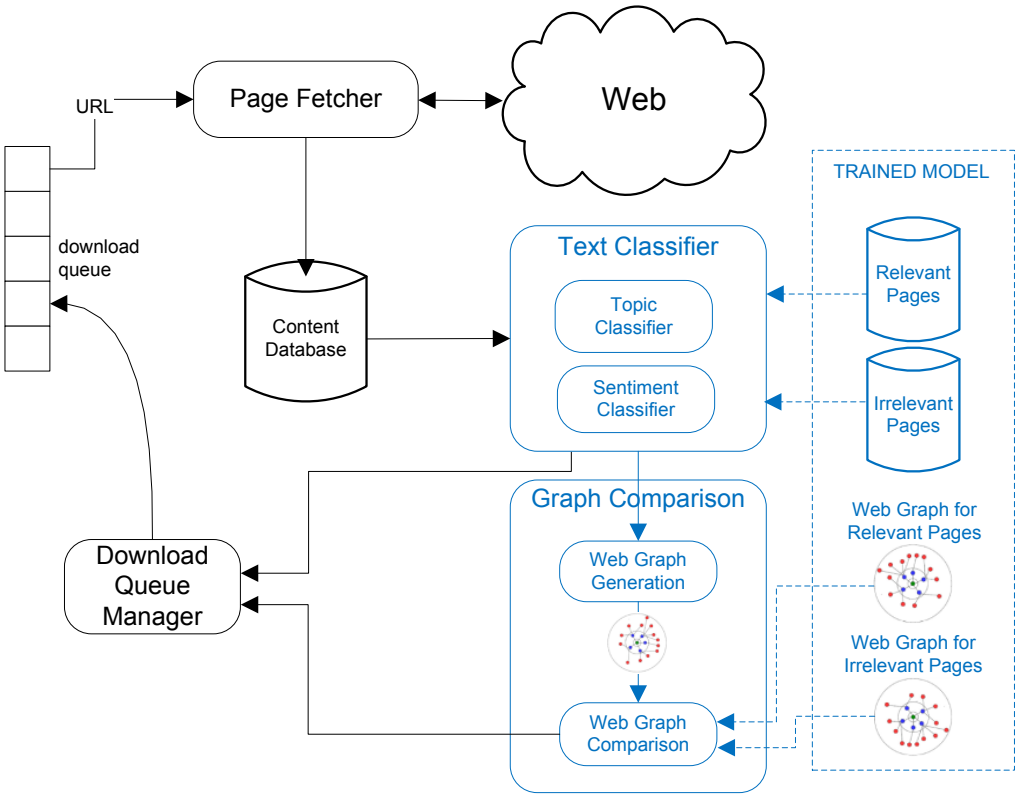


Figure 2.4: Web crawling architecture of graph-based sentiment crawler by Fu et al. [42]

basic focused web crawling architecture in Figure 2.2, as shown in Figure 2.4. As illustrated, graph-based sentiment crawler consists of four components:

- **Crawler**: This is the page fetcher component of the basic web crawling architecture. It basically picks the top URL from the download and crawls it.
- **Text classifier**: Text classifier consists of a topic classifier for computing the relevance of a crawled page using a trained model and a sentiment classifier for estimating a sentiment score to a crawled page.

  Topic classifier heavily depends on the trained model, which weights the keywords in relevant and irrelevant pages of training dataset by using the information gain heuristic. Then, topic classifier computes the sum of weights of each keyword in the page based on the occurrence distribution across classes. When the crawled page has a positive topic relevance score, it is considered as topic relevant.

  Sentiment classifier computes a sentiment score by considering only the semantic weight of sentences containing relevant keywords according to the trained model. For semantic weights, this classifier utilizes SentiWordNet [39], which is a lexical resource having more than 150,000 words with three sentiment polarity score (positivity, negativity, objectivity). When the sentiment score of the crawled

page differs from the relevant pages by less than a threshold, it is considered as sentiment relevant.

Using the scores of the classifiers, text classifier component categorizes a crawled page as one of the classes; $C1$ for relevant topic and sentiment, $C2$ for relevant topic only, $C3$ for relevant sentiment only, and $C4$ for irrelevant topic and sentiment. Out-going links of $C1$ pages are forwarded to queue manager with highest weights. When a page is categorized as other than $C1$, the weights of out-going links are calculated by graph comparison component using web graphs.

- **Graph comparison**: This component simply determines for "tunelling" by analysing the irrelevant pages whether they are likely to lead to relevant pages by applying graph comparison. In other words, this component calculates the similarity of discovered web graph with web graphs of trained data.
- **Queue manager**: This is the download queue manager component of the basic web crawling architecture. It mainly ranks the URLs in descending order based on their weights determined by text classifier and graph comparison components.

In empirical evaluations, the graph-based sentiment crawler is compared against traditional topic-driven crawlers, including Vector Space Model, Keyword-based method, Context Graph Model, Hopfield Net, PageRank and Breadth-First Search. Two test beds are used for evaluation as animal rights content for corporate social responsibility (with 525K web pages) and medicine content for post-marketing drug surveillance (with 12.3M web pages). According to the reported experiments, the graph-based sentiment crawler outperforms other baseline crawling techniques in terms of the F-measure, precision, and recall metrics on two different test beds.

2.4.2 Emotional Search Engine

Recently, there has been a number of initiatives for the sentimental and emotional search engine in the literature. *Opinion Crawl*[5] and *We Feel Fine*[6] are the very first initiatives.

Opinion Crawl is a commercial engine having a standard textual interface for searching sentimental content on a subject (e.g. a person, an event, a company or product). It is reported that the crawler of the system searches the latest pages on many popular subjects or current public issues, and calculates sentiment for them on an ongoing basis [1]. Therefore, the crawler neither predicts the sentimentality of an "unseen" page, nor ranks the entries considering their potential for sentimentality.

We Feel Fine [55] is an emotional search engine with a web-based data visualization capability to explore and analyse people's emotions qualitatively. The only source

[5] Opinion Crawl, `http://www.opinioncrawl.com/`
[6] We Feel Fine, `http://wefeelfine.org`

of the system is the weblogs, including microblogs and social networking sites. The system continuously crawls the newly posted blog entries and harvests human feelings by checking occurrences of the phrases "I feel" and "I am feeling" only. When the system detects such a phrase, it checks if it includes one of about 5,000 pre-identified feelings, which consist of adjectives and some adverbs. If a valid feeling is found, the system indexes the sentence together with the feeling and author's demographics including age, gender and location. Kamvar and Harris [55] report that the system with a database of over 14 million feelings, is capable of performing searches on feelings by using parameters of feeling, age, gender, location, weather and date. Although We Feel Fine is an emotional search engine, it does not use the notion of ranking stating that it is difficult and unreasonable to rank sentiment. Additionally, the crawler component of the system is not sentiment-focused. The crawler component, regularly fetching new blog entries from a certain list of blog pages, does not ranks the "unseen" pages in the download queue with respect to their potential for sentimentality.

CHAPTER 3

SENTIMENT-FOCUSED WEB CRAWLING

"You never change things by fighting the existing reality. To change something, build a new model that makes the existing model obsolete."

– Buckminster Fuller

The timely discovery of sentimental content in the web is important as most sentiments or opinions quickly loose their value if they are not immediately discovered. Interestingly, so far, the discovery of such content has not received much attention from the research community. In this chapter, we propose a sentiment-focused web crawling framework to facilitate the quick discovery of sentimental content. Within this framework, we propose different techniques to estimate the sentimentality of an "unseen" web page and guide the crawling process through these estimates.

The chapter is organized as follows. We introduce sentiment-focused web crawling problem in Section 3.1. In Section 3.2, we propose our framework for sentiment-focused web crawling. Section 3.3 presents the results of the user study conducted in order to create a ground-truth for sentimentality. In Section 3.4, we present the techniques to predict the sentimentality of an "unseen" web page. We provide details on dataset characteristics and experimental setup in Section 3.5. The experimental results are given in Section 3.6. In Section 3.7, we conclude with discussions on our experimental results together with some future work, and compare our framework with the prior work.

3.1 Problem

Let us assume that we have a set $\mathcal{P}$ of N web pages to be crawled. Let us denote the set of outgoing links of a page $p_j \in \mathcal{P}$ by $\mathcal{L}_j$. At any point in time during the crawling, a page can belong to one of the following three sets: crawled ($\mathcal{C}$), discovered but not yet crawled ($\mathcal{D}$), or undiscovered ($\mathcal{U}$). We represent the sets obtained after crawling i pages by $\mathcal{C}_i$, $\mathcal{D}_i$, and $\mathcal{U}_i$, respectively. We assume that each page $p_j \in \mathcal{C}_i$

is associated with a sentiment score s_j. Before the crawling starts, we have $\mathcal{C}_0 = \{\}$, $\mathcal{D}_0 = \mathcal{S}$, and $\mathcal{U}_0 = \mathcal{P} - \mathcal{S}$, where $\mathcal{S}$ is a set of seed pages selected from $\mathcal{C}$. When the crawling terminates, we have $\mathcal{D}_{|\mathcal{C}|} = \{\}$ and $\mathcal{C}_{|\mathcal{C}|} \cup \mathcal{U}_{|\mathcal{C}|} = \mathcal{P}$. Note that $\mathcal{C}_{|\mathcal{C}|} = \mathcal{P}$ does not necessarily hold as some pages in $\mathcal{P}$ may not be reachable by the crawler. Crawling a page $p_j \in \mathcal{D}$ at iteration i leads to $\mathcal{C}_i = \mathcal{C}_{i-1} \cup \{p_j\}$, $\mathcal{D}_i = (\mathcal{D}_{i-1} \cup \mathcal{L}_j) - \{p_j\}$, and $\mathcal{U}_i = \mathcal{U}_{i-1} - \mathcal{L}_j$.

In the sentiment-focused web crawling problem, our objective is to find a sequence $\langle p_{j_1}, \ldots, p_{j_i}, \ldots, p_{j_n} \rangle$ of n pages such that the total sentimentality $S(\mathcal{C}_n)$ of the pages in $\mathcal{C}_n$ is maximized, i.e., we would like to maximize

$$S(\mathcal{C}_n) = \sum_{i=1}^{n} s_{j_i}. \tag{3.1}$$

Here, n is a parameter such that $n \ll N$, i.e., the goal is to maximize the total sentimentality of the crawled pages at the early stages of the crawling.

3.2 Framework

3.2.1 Architecture

The high-level architecture of sentiment focused crawler is depicted in Figure 3.1. As illustrated, sentiment focused crawler is composed of following main parts:

- **Page retrieval component**: This component fetches the web pages from the Web by picking up the URLs from the download queue.
- **Storage component**: This component is responsible for storing the retrieved web pages in the Content Database.
- **Text processing component**: This component processes each retrieved web page to extract features. First, the textual content is extracted by parsing and cleaning. We investigate different parser alternatives for this purpose. The obtained textual content is then tokenized into sentences and words. Second, the URLs that are linked by the page are extracted. Finally, features associated with the textual content, URLs, and anchor text, are extracted.
- **Prediction component**: This component is composed of two sub-modules: score estimator and download queue manager. The main task of score estimator is to predict the sentimentality (or polarity) for extracted URLs using the features provided by the feature extractor. After assignment of sentiment (or polarity) scores, extracted links are given to the download queue manager, which reorganizes the URLs in the download queue in decreasing order of the predicted

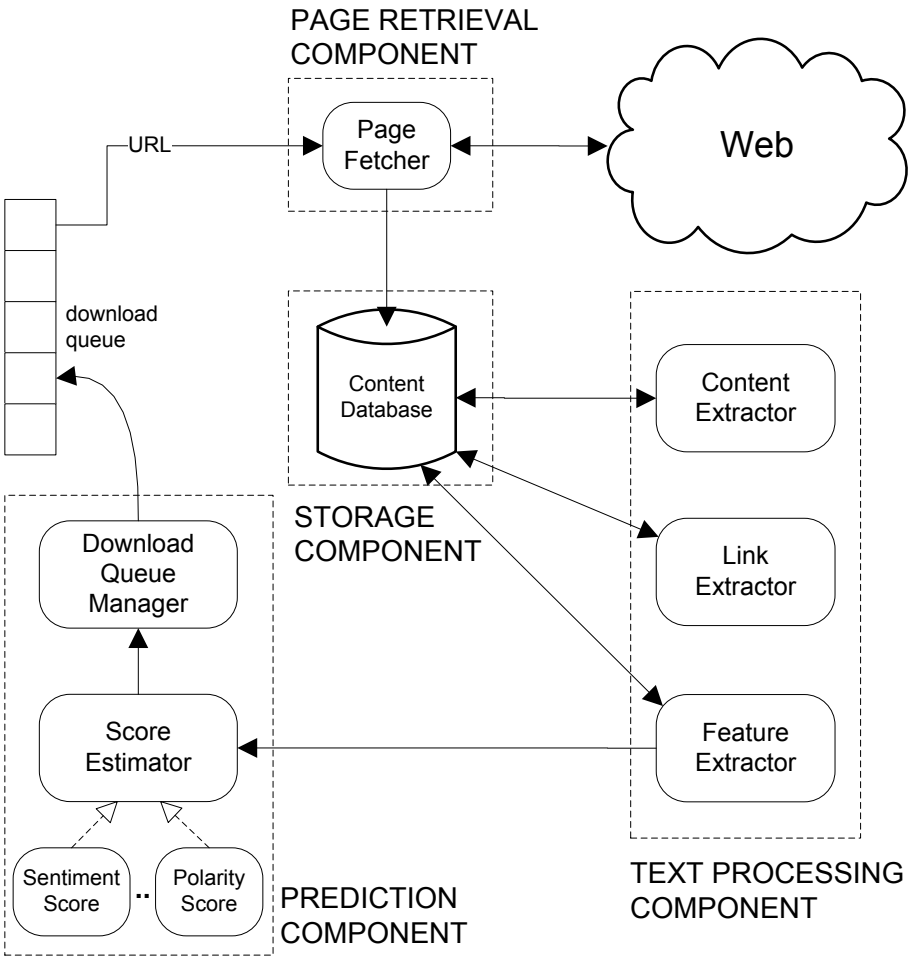


Figure 3.1: Generic architecture for sentiment-focused crawling.

scores. The URL with the highest sentiment (or polarity) score is passed to the page fetcher as the next URL to be downloaded.

In this framework, some issues, such as DNS caching, politeness policies, refresh rates, performance monitoring, handling URLs with particular file extensions that are not of interest, and handling hidden Web are ignored as they are not essential for our purpose. Those features can be handled as in a general purpose crawler.

3.2.2 Tools

The proposed sentiment-focused web crawler design requires a number of important components, for content and sentence extraction from web page in "Page processing module", and sentiment analysis in "Sentiment-focused crawling module".

- **Content extraction**: Content extraction is a process of removing HTML tags and cleaning the textual content to retrieve the body content of the web page. In the cleaning part, for the proper sentiment analysis, items such as advertisements, headers, footers and menu bars should be excluded from the main content. For such a cleaning and content extraction phase, we analyze two alternative tools: HTML Parser[1] and BoilerPipe[2]. Despite its simplicity and having well documented API, HTML Parser software performs relatively poor in cleaning the content since it extracts almost all text found in a web page. The second alternative, BoilerPipe is a specific tool for extracting the "body" content of web pages. Out of five extraction options provided by BoilerPipe, "CanolaExtractor" is evaluated to be the best option since it obtained the best extraction performance over a larger number of pages. Sample text extraction of a blog page performed by "CanolaExtractor" of BoilerPipe is shown in Figure 3.2.

- **Sentence and word extraction**: Stanford Core NLP library[3] is used for tokenizing and splitting of sentences and words from the extracted page content. Stanford Core NLP is a powerful suite of natural language analysis tools which can give the base forms of words, their parts of speech and mark up the structure of sentences from an input of raw English text.

- **Sentiment analysis**: As stated in the proposed framework, a sentiment analysis tool is required for analysing several features, such as page content and anchor text. For this purpose, a lexicon-based sentiment analysis tool, SentiStrength software [83] is used (see Section 2.3.4 for details).

  As an alternative to original lexicon file, we have proposed to use subset of the lexicon that contains only the emotional adjectives and adverbs. This subset includes 444 entries in total.

  Though we utilize the SentiStrength tool, we have an important issue of how to quantify the sentimentality of a web page content. A web page can be described as a "set of sentences", therefore sentimentality of a page can be computed by use of the sentiment analysis of each sentences. The page sentimentality score can be formulated by two alternatives.

  – The average of sentimentality scores of sentences (SS): This sentimentality metric is first formulated by Kucuktunc et al. in [60]. This metric is defined as follows: let $\mathcal{S}_i$ be the set of sentences in a page p for which we want to computer the sentiment score. The sentiment score of p is estimated as the average of sentence sentiment scores:

$$SS_p = \frac{1}{i} \times \sum_{\mathcal{S}_i \in p} ({s_i}^+ - {s_i}^- - 2). \tag{3.2}$$

[1] HTML Parser (version 2.0), `http://htmlparser.sourceforge.net`.
[2] BoilerPipe (version 1.2.0), `http://code.google.com/p/boilerpipe`.
[3] Stanford Core NLP library (version 1.2.0), `http://nlp.stanford.edu/software/corenlp.shtml`.

Figure 3.2: A text exraction sample by BoilerPipe.

– The average over the absolute values of all individual words in the page (*WS*): We propose to use this metric and in our experiments, we have seen that it is the best performing technique in polarity detection of Turkish texts [88]. The sentiment score of a page p is estimated as the average of absolute values of words in all sentences:

$$WS_p = \frac{1}{|word|} \times \sum_{word_k \in \mathcal{S}_i} |s_k|. \tag{3.3}$$

- **Machine learning model**: The sentiment-focused crawler we propose, relies on machine learning and builds a predictive model using certain extracted features of crawled pages. LibSVM software[4] developed by Chang and Lin [28], is used with the regression mode for this learner.

3.3 User Study

As stated in [87], evaluating the performance of different focused crawling techniques in fetching sentimental content requires knowing the actual sentimentality scores of fetched pages. Unfortunately, there has not been any study on sentiment score assignment to a web page collection. Therefore, we need to create a ground-truth for sentiment scores either by forming an editorial team to label the pages or by using some score estimates that substitute the actual sentiment scores. In this study, second alternative is chosen due to large size of the web page collection.

As explained in Chapter 3.2.2 and Chapter 3.4, there are a couple of alternatives for tools in the framework and methods for sentimentality prediction. By the help of this user study, we aim to identify the best combination of these alternatives that yield the most accurate page sentiment scores. The alternatives are as *i)* textual content extraction using HTML Parser (`HP`) or BoilerPipe (`BP`), *ii)* sentiment score computation based on the sentence scores (`SS`) or word scores (`WS`), and *iii)* lexicon for SentiStrength as all sentimental words (`All`) or only the sentimental adjectives (`Adj`). This yields to eight possible parameter combinations to be considered as summarized in Table 3.1.

A small-scale user study with 5 judges (`J1`, `J2`, `J3`, `J4`, and `J5`) is conducted to identify the best performing parameter combination listed in in Table 3.1. For the user study, a sample of 500 web pages is randomly chosen from the ClueWeb09[5] web page collection. The judges individually evaluate and assign the labels "sentimental" or "not sentimental" to a page.

[4] LibSVM (version 3.1.12), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
[5] ClueWeb09 – `http://lemurproject.org/clueweb09.php`

Table3.1: Parameter combinations for sentiment-focused crawling framework

	Text Extraction	Sentiment Score	Lexicon
`HP-SS-All`	HTML Parser	sentence scores	all words
`HP-SS-Adj`	HTML Parser	sentence scores	sentimental adjectives
`HP-WS-All`	HTML Parser	word scores	all words
`HP-WS-Adj`	HTML Parser	word scores	sentimental adjectives
`BP-SS-All`	BoilerPipe	sentence scores	all words
`BP-SS-Adj`	BoilerPipe	sentence scores	sentimental adjectives
`BP-WS-All`	BoilerPipe	word scores	all words
`BP-WS-Adj`	BoilerPipe	word scores	sentimental adjectives

Based on the labeling of the judges, following three separate ground-truths are defined for the judged pages:

- **Ground Truth-1** (`GT1`): In this scenario, a page is assumed to be sentimental if at least one judge thinks so.
- **Ground Truth-2** (`GT2`): This scenario works according to majority voting. A page is assumed to be sentimental when the majority of judges think so.
- **Ground Truth-3** (`GT3`): This scenario is the most strict one. A page is labeled as sentimental only if all the judges agree on the sentimentality of the page.

Overlap metric and Cohen's kappa [25] are used for evaluating the agreement between different judges. The overlap metric simply measures the rate at which two judges agree in the labels they assigned to a page. Cohen's kappa (Equation 3.4) takes into account the likelihood of agreement by chance. In Equation 3.4, *Pr(a)* is the relative observed agreement among judges, and *Pr(e)* is the hypothetical probability of chance agreement. When the judges are in complete agreement, kappa value (κ) returns 1, otherwise 0.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{3.4}$$

Table 3.2 and Table 3.3 display the agreement between judges as well as their agreement with the ground-truths. The fraction of sentimental pages is identified as 0.23, 0.24, 0.18, 0.23 and 0.33 by the judges (the S column in the table), who labelled about one-fourth of the pages as sentimental, on average. According to the table, the overlap (O) between judges is above 85 %, which shows a high agreement between the judges. The kappa values (κ) also support this inference.

Table3.2: The degree of agreement among the judges in terms of overlapping

		Overlap (O)							
Judge	S	J1	J2	J3	J4	J5	GT1	GT2	GT3
J1	0.23	1.00	0.87	0.87	0.87	0.86	0.77	0.93	0.85
J2	0.24	0.87	1.00	0.85	0.87	0.77	0.77	0.91	0.84
J3	0.18	0.87	0.85	1.00	0.84	0.74	0.71	0.90	0.91
J4	0.23	0.87	0.87	0.84	1.00	0.77	0.76	0.93	0.85
J5	0.33	0.86	0.77	0.74	0.77	1.00	0.87	0.82	0.75
Avg.	0.24	0.89	0.87	0.86	0.87	0.83	0.78	0.90	0.84

Table3.3: The degree of agreement among the judges in terms of Kappa values

		Kappa (κ)							
Judge	S	J1	J2	J3	J4	J5	GT1	GT2	GT3
J1	0.23	1.00	0.64	0.61	0.62	0.66	0.51	0.79	0.46
J2	0.24	0.64	1.00	0.55	0.64	0.44	0.53	0.77	0.44
J3	0.18	0.61	0.55	1.00	0.51	0.33	0.40	0.67	0.59
J4	0.23	0.62	0.64	0.51	1.00	0.44	0.51	0.79	0.46
J5	0.33	0.66	0.44	0.33	0.44	1.00	0.73	0.56	0.30
Avg.	0.24	0.71	0.65	0.60	0.64	0.57	0.54	0.72	0.45

As a result of the high inter-judge agreement, it can be concluded that the labels obtained through the user study form a sufficiently reliable basis to evaluate the performance of the parameter combinations. In the next step of the user study, for each parameter combinations we compute the sentiment scores for our sample pages and obtain 8 different rankings. As the evaluation metrics for the performance of these rankings, the average precision (AP) metric, discounted cumulative gain (DCG) metric, and the precision values obtained at ranks 10, 50, and 100, are used. In addition, a random ranking where the pages are randomly sorted, is defined as the baseline. In this case, the metrics reflect the average of one million trials, each started with a different random seed.

Precision, average precision and discounted cumulative gain metrics are popular performance and correctness measures for information retrieval systems [65]. Precision, formulated in Equation 3.5, can be described as the fraction of the documents retrieved that are relevant to the user's information need.

$$precision = \frac{|\{relevant\ documents\} \bigcap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{3.5}$$

Average precision gives a performance metric while considering the order in which the returned documents are presented. Equation 3.6 formulates the average precision, where *P(k)* is the precision at cut-off k and *rel(k)* is the binary function showing whether the item at rank k is relevant.

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{number\ of\ relevant\ documents} \tag{3.6}$$

Finally, discounted cumulative gain measures the usefulness of a document based on its rank in the result list using a graded relevance scale. The discounted cumulative gain accumulated at a particular rank position p is defined as shown in Equation 3.7.

$$DCG_p = rel_1 + \sum_{i=1}^{p} \frac{rel_i}{log_2 i} \tag{3.7}$$

Table 3.4 provides the computed metrics for different rankings, each generated using a different parameter combination. In the table, **bold** values correspond to the best results for the related metric. As seen, all of the rankings perform considerably better than the baseline. The `BP-WS-Adj` combination yields the most accurate rankings for the relatively relaxed `GT1` and the more conservative `GT3` scenarios. Interestingly, `HP-WS-Adj` combination performs slightly better than `BP-WS-Adj` combination for the `GT2` scenario. For `GT1` scenario, most of the retrieved pages are sentimental pages whereas the performance drops dramatically for the `GT3` scenario hence the problem becomes more difficult.

Table 3.5 summarizes the performance of individual parameter alternatives (e.g., the metrics reported for `HP` are averages over `HP-SS-All`, `HP-SS-Adj`, `HP-WS-All`, and `HP-WS-Adj`). According to the table, it is interesting to note that all three types of parameters have a winner. BoilerPipe performs slightly better than HTML Parser. The reason behind this can be explained by the better capability on HTML content cleaning. Another finding is that the lexicon containing adjectives yields a performance better than using the default SentiStrength lexicon. It can be stated that the default SentiStrength lexicon has superfluous amount of sentimental words, some of which are context-dependant. Finally, the word-level sentiment scores provides significant improvement rather than the sentence-level scores. This result indicates that a fine-grain (at the word level) aggregation of the sentiment scores is more promising for evaluating sentimentality.

As a result of all these findings from Table 3.4 and Table 3.5, it can be concluded that `BP-WS-Adj` is the best alternative for the ground-truth for our web page collection.

Table3.4: The ranking quality achieved by different parameter combinations over 500 randomly sampled pages

GT	Metric	Rand	HP-SS-All	HP-WS-All	BP-SS-All	BP-WS-All	HP-SS-Adj	HP-WS-Adj	BP-SS-Adj	BP-WS-Adj
	P@10	0.46	0.60	**1.00**	0.80	**1.00**	0.50	**1.00**	**1.00**	**1.00**
	P@50	0.46	0.54	0.78	0.78	0.78	0.70	0.82	0.86	**0.92**
GT1	P@100	0.46	0.68	0.72	0.76	0.75	0.68	0.77	0.81	**0.82**
	AP	0.47	0.63	0.69	0.71	0.72	0.65	0.71	**0.74**	**0.74**
	DCG	32.74	34.77	37.24	36.61	37.53	35.95	37.46	37.63	**37.82**
	P@10	0.22	0.50	0.70	0.70	0.80	0.50	**1.00**	0.50	0.90
	P@50	0.22	0.42	0.60	0.52	0.54	0.46	**0.64**	0.50	0.60
GT2	P@100	0.22	0.44	0.49	0.47	**0.52**	0.43	0.49	0.50	**0.52**
	AP	0.23	0.44	0.51	0.51	**0.55**	0.45	**0.55**	0.47	**0.55**
	DCG	15.52	18.09	20.25	19.58	20.54	19.20	**20.71**	19.38	20.67
	P@10	0.08	0.20	0.20	0.10	0.20	0.40	**0.60**	0.30	0.40
	P@50	0.08	0.20	0.32	0.22	0.28	0.24	0.32	0.28	**0.38**
GT3	P@100	0.08	0.21	0.23	0.22	**0.26**	0.23	0.25	**0.26**	**0.26**
	AP	0.09	0.22	0.24	0.22	0.26	0.29	0.37	0.30	**0.40**
	DCG	5.79	7.27	7.43	7.06	7.63	8.67	9.07	8.67	**9.42**

Table3.5: The ranking quality observed for individual parameter alternatives

GT	Metric	HP	BP	SS	WS	All	Adj
	P@10	0.78	0.95	0.73	1.00	0.85	0.88
	P@50	0.71	0.84	0.72	0.83	0.72	0.83
GT1	P@100	0.71	0.79	0.73	0.77	0.73	0.77
	AP	0.67	0.73	0.68	0.72	0.69	0.71
	DCG	36.36	37.40	36.24	37.51	36.54	37.22
	P@10	0.68	0.73	0.55	0.85	0.68	0.73
	P@50	0.53	0.54	0.48	0.60	0.52	0.55
GT2	P@100	0.46	0.50	0.46	0.51	0.48	0.49
	AP	0.49	0.52	0.47	0.54	0.50	0.50
	DCG	19.56	20.04	19.06	20.54	19.62	19.99
	P@10	0.35	0.25	0.25	0.35	0.18	0.43
	P@50	0.27	0.29	0.24	0.33	0.26	0.31
GT3	P@100	0.23	0.25	0.23	0.25	0.23	0.25
	AP	0.28	0.30	0.26	0.32	0.24	0.34
	DCG	8.11	8.19	7.92	8.39	7.35	8.96

3.4 Sentimentality Prediction

According to Pant et al. in [74], link content (e.g., referrer page), ancestor pages (e.g., pages that lead to the referrer page), and in general web graphs have been used by focused web crawlers for estimating the benefit of following a particular URL. In accordance with this statement, we claim that the sentimentality of an "unseen" page can be predicted by using different features of the referrer pages together with the known features of target page.

Using the previously introduced notation in Section 3.1, let $\mathcal{C}_i$ denote the set of already downloaded pages after crawling i pages, and let $\mathcal{D}_i$ denote the set of pages whose URLs are discovered, but content is not yet crawled. If we assume that $\mathcal{I}_{i,j}$ refers to the set of currently crawled pages that provide a link to page p_j, i.e., $\mathcal{I}_{i,j} \subset \mathcal{C}_i$, $p_k \in \mathcal{I}_{i,j}$, $p_j \in \mathcal{D}_i$, and there exists a link $(p_k \rightarrow p_j)$ with anchor text $anchor_{kj}$, then our problem can be formulated as estimating sentiment score s'_j using the certain features of pages in $\mathcal{I}_{i,j}$.

We propose three alternative techniques for sentimentality prediction of "unseen" pages using certain features extracted from the previously discovered pages and links referring to the page:

- **Based on referring anchor text**: This technique simply aggregates the actual sentiment scores of referring anchor texts. After i pages are crawled, the sentiment score s'_j of page p_j is simply estimated as an average of the actual sentiment scores of anchor texts in all referring pages in $\mathcal{I}_{i,j}$:

$$s'_j = \frac{1}{|\mathcal{I}_{i,j}|} \times \sum_{p_k \in \mathcal{I}_{i,j}} s_{anchor_{kj}}. \tag{3.8}$$

- **Based on referring page content**: This technique simply aggregates the actual sentiment scores of referring pages. After i pages are crawled, the sentiment score s'_j of page p_j is simply estimated as an average of the actual sentiment scores of all referring pages in $\mathcal{I}_{i,j}$:

$$s'_j = \frac{1}{|\mathcal{I}_{i,j}|} \times \sum_{p_k \in \mathcal{I}_{i,j}} s_k. \tag{3.9}$$

- **Based on machine learning**: In this approach, we build a machine learning model M over the previously downloaded web pages in $\mathcal{C}_i$. An instance I_j in the model M refers to a target web page. The prediction target is the actual sentiment score s_j of the page p_j. An instance I_j involves n different features $\langle f_{j_1}, f_{j_2}, \ldots, f_{j_n} \rangle$, including the above-mentioned average sentiment scores computed over the referring pages and anchor text as well as some other statistics.

Table3.6: Features used by the learning model

Type	Feature description
Referring page	Average page size (w/ HTML tags)
	Average page size (w/o HTML tags)
	Average number of DOM objects
	Average number of pictures in the page
	Average number of out-going links
	Average number of sentences
	Average number of words
	Average number of unique words
	Average length of a sentence
	Average number of self links
	Ratio of number of links to page size
	Ratio of HTML portion of links to page size HTML
	Average sentiment score of sentences
	Average sentiment score of words
	Average sentiment score of keywords in meta part
	Average sentiment score of page titles
	Maximum sentiment score of sentences
	σ of sentence sentiment scores
Anchor text	Average number of terms in referring anchor text
	Average sentiment score of referring anchor text
Page URL	Sentiment score for the entire URL
	Sentiment score for the host part of the URL

The complete list of features is given in Table 3.6. We note that the sentiment scores can be more accurately estimated as more referring pages become available for the target page, i.e., set $\mathcal{I}_{i,j}$ gets larger. Therefore, the model is rebuild at regular intervals during the crawling using the pages downloaded so far and is used to predict the sentiment scores of the pages in set $\mathcal{D}_i$.

3.5 Experimental Setup

3.5.1 Dataset

In our experiments, we need a huge web crawl dataset that is not specific to any subject. Due to the dynamic nature of the Web [12, 30], using a web crawl dataset (instead of real-time crawling) is thought to be beneficial for us since all the experiments will be reproducible, which cannot be possible on continuously changing web data. In the literature, ClueWeb09 is the most popular dataset in academic research. This dataset, collected between January and February of 2009, includes 1 billion web pages with content (25 TB size).

ClueWeb09 dataset serves only the web page archive with the corresponding web address. In addition, there are some publicly available derived data, such as PageRank and spam scores. PageRank scores are computed by Carnegie Mellon University[6], applying the basic PageRank algorithm with random start probability 0.15. The spam scores are provided by Waterloo Spam Rankings[7] computed by using a simple content-based classifier [33]. The spam scores are within a range of 0 to 100 and it is suggested by the creators that pages having a spam score less than 70 are assumed as spam. There are four sets of spam scores in Waterloo Spam Rankings and in this work "Fusion" subset is used as suggested by the creators of Waterloo Spam Rankings.

For the experiments, we use the TREC Category-B part of ClueWeb09 collection, which is a subset of archive including the first 50 million English pages. We remove Wikipedia pages from the collection as those pages are mainly informational and do not include much sentiments, which leaves us with a large sample containing 44,218,678 web pages (1.2 TB on disk). The dataset has the following statistics:

- A page contains 669.8 words (231.3 unique words) and 14.1 sentences, on average.
- A page has 55.5 outgoing and 20.7 incoming links on average.
- A page has a size of 28,029 and 3,037 bytes before and after parsing the HTML tags, respectively.
- About 8.33% do not link to any other page whereas about 16.2% do not receive a link from any other page.
- About 47.4% fall into the spam category.
- Majority of the pages belongs to `com` domain with rate 74.5%. The other significant domains for pages are `org` (12.0%), `edu` (5.6%), `net` (4.8%), `gov` (1.9%), and `info` (1.2%). Some domain distribution statistics are given in Table 3.7.
- Similar to page domains, majority of the outgoing links belong to `com` domain with rate 79.2%. The other significant domains for links are `org` (8.7%), `net` (6.3%), `edu` (3.5%), `gov` (1.0%), and `info` (0.9%).
- Sentiment score, spam score and PageRank distributions for full experiment dataset are shown in Table 3.8.

Figure 3.3 shows the distribution of the above-mentioned scores with respect to the rank of a page in the distribution. The values in each curve are scaled to [0,1] range. In terms of skewness, the polarity and PageRank scores have the most skewed distribution, followed by the sentiment score distribution. The spam scores follow a more regular distribution as they are obtained by creating equal-size score bins over another

[6] `http://boston.lti.cs.cmu.edu/clueweb09/wiki/tiki-index.php?page=PageRank`
[7] `http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/`

Table3.7: Some domain distribution statistics for experiment dataset

Domain	%	Sentiment Score (Avg.)	Polarity Score (Avg.)	PageRank Score (Avg.)	Spam Score (Avg.)
aero	0.004	0.0015	0.0011	0.1803	83.99
asia	0.010	0.0029	0.0019	0.1636	39.59
biz	0.213	0.0040	0.0032	0.1791	46.88
cat	0.003	0.0013	0.0010	0.1640	69.89
com	74.480	0.0044	0.0032	0.1808	61.64
coop	0.012	0.0038	0.0032	0.1895	85.05
edu	5.609	0.0016	0.0010	0.1858	85.80
gov	1.894	0.0011	0.0005	0.1915	86.02
info	0.761	0.0033	0.0023	0.1703	37.59
int	0.056	0.0007	0.0004	0.1780	85.18
jobs	0.001	0.0138	0.0137	0.1807	89.44
mil	0.079	0.0010	0.0005	0.1911	83.88
mobi	0.012	0.0039	0.0030	0.1714	40.73
museum	0.005	0.0015	0.0009	0.2060	84.32
name	0.020	0.0041	0.0027	0.1689	44.80
net	4.821	0.0042	0.0027	0.1789	57.93
org	11.996	0.0028	0.0017	0.1837	77.56
pro	0.005	0.0045	0.0039	0.1729	41.78
travel	0.020	0.0048	0.0043	0.1773	74.71

Table3.8: Some distribution statistics for experiment dataset

Distribution	Min.	Avg.	Max.	Std. dev.
Sentiment score	0.000	0.004	1.781	0.009
Polarity score	−1.000	0.003	1.600	0.008
Spam score	0.000	65.000	99.000	27.231
PageRank	0.150	0.181	1,080.164	0.421

score distribution. Figure 3.4 illustrates that there is no correlation between sentiment and polarity scores of a page. The scatter plots in Figures 3.5- 3.6 and Figures 3.7-3.8 show the spam score and PageRank distributions with respect to the sentiment score and polarity score (the sample is down-sized a thousand times), respectively. In general, we observe no clear correlation of the spam score to both sentiment and polarity scores, i.e., the likelihood of a page to contain spam and its sentimentality or polarity seem to be pretty much independent. On the other hand, there is some weak correlation of the PageRank values to the sentiment and polarity scores of pages since higher PageRank values are attained only by the pages having less sentimental and neutral content.

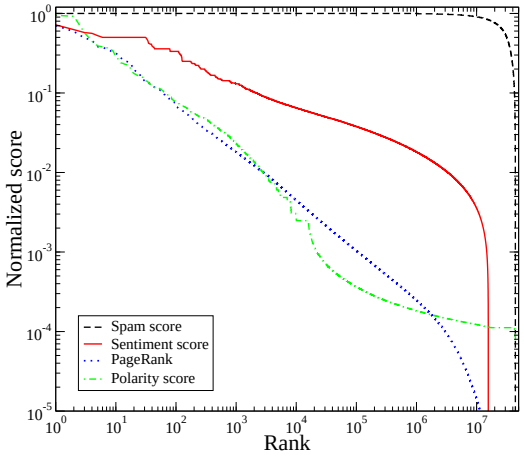


Figure 3.3: Normalized sentiment, polarity, spam, and PageRank score distributions (log-log scale).

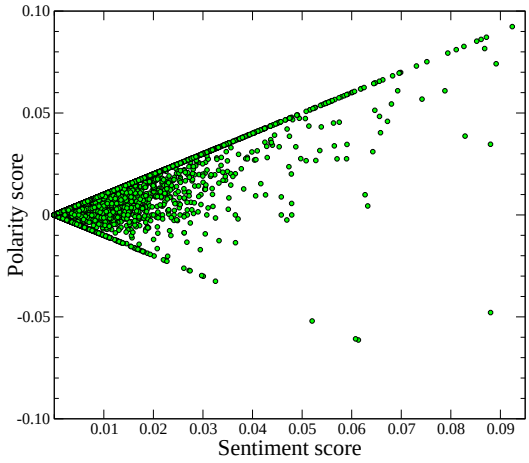


Figure 3.4: Sentiment score versus polarity score.

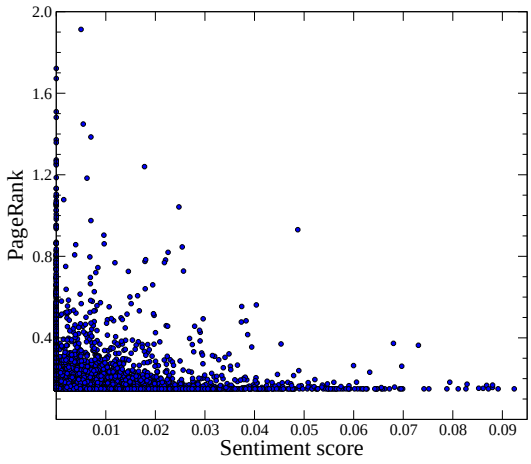


Figure 3.5: Sentiment score versus PageRank.

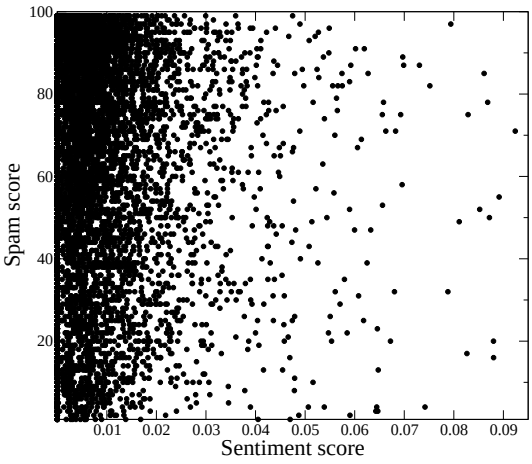


Figure 3.6: Sentiment score versus spam score.

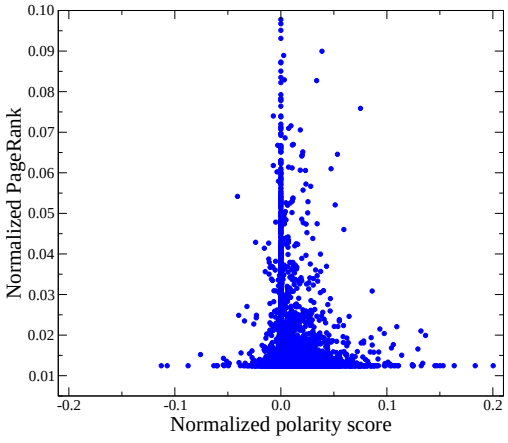


Figure 3.7: Polarity score versus PageRank.

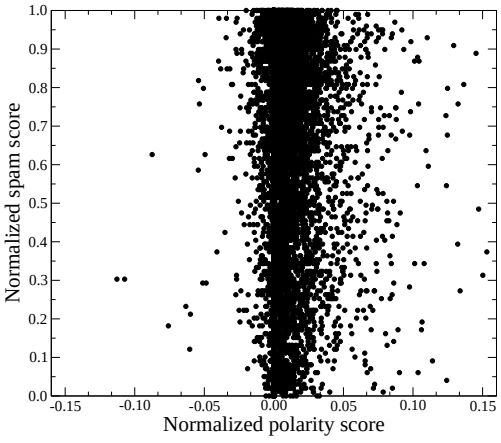


Figure 3.8: Polarity score versus spam score.

3.5.2 Setup

For the experiments, the framework given in Section 3.2 is simulated with an implementation in Java 6.0[8]. MySQL[9] (version 5.1.61) database is used for storing the web pages and the extracted features. Experiments are executed on a 16-core computer with 48GB of RAM running Debian Linux.

3.5.3 Crawlers

Three sentiment-focused crawlers, each adopting the sentimentality prediction techniques described in Section 3.4, has been implemented and evaluated against baseline and oracle crawlers. Our expectation for sentiment-focused crawlers is to outperform baseline crawlers while approaching to theoretical sentiment-focused crawler.

- **Sentiment-focused crawlers**
  - **Anchor text based** (`P-AT`) : This is the sentiment-focused crawler, which predicts the sentiment score of a target page by aggregating the actual sentiment scores of referring anchor texts.
  - **Page content based** (`P-PC`) : This is the sentiment-focused crawler, which predicts the sentiment score of a target page by aggregating the actual sentiment scores of referring pages.
  - **Machine Learning based** (`P-ML`) : This is the sentiment-focused crawler based on machine learning, in which the sentiment score of a target page is predicted by a machine learning model using the features given in Table 3.6.
- **Baseline crawlers**
  - **Random crawler** (`B-RA`) : This is a general-purpose focused crawler, in which each time page retrieval module randomly picks a URL from the download queue.
  - **Highest indegree crawler** (`B-ID`) : This is a general-purpose focused crawler, that prioritizes the URLs according to their indegree value, which is the total number of referring pages fetched so far.
  - **Lowest depth (BFS) crawler** (`B-BF`) : This is a general-purpose focused crawler that works in breadth-first search order.
- **Oracle crawlers**
  - **Theoretical sentiment-focused crawler** (`O-SE`) : This oracle crawler has perfect knowledge of the individual scores of pages so that it prioritizes

[8] Java, `http://www.oracle.com/technetwork/java/`
[9] MySQL, `http://www.mysql.com/`

Table3.9: Coverage of pages with different sizes and types of seed pages

	Random		Highest Outgoing Link	
Seed Size	With spam filtering	Without spam filtering	With spam filtering	Without spam filtering
1	8,760,536	19,848,657	8,760,533	19,847,864
10	8,760,542	19,848,794	8,760,536	19,847,865
20	8,760,555	19,848,894	8,760,536	19,847,871
50	8,760,847	19,874,890	8,760,536	19,849,271
100	8,760,991	19,878,727	8,760,536	19,850,348
200	8,761,407	19,886,745	8,760,933	19,853,832

the URLs according to their actual sentiment scores. Therefore it forms the upper bound for a crawling process and indicates how much the other types of crawlers can improve.

- **Spam Score crawler** (`O-SP`) : This oracle crawler prioritizes the URLs according to spam scores from Waterloo Spam Rankings.
- **PageRank crawler** (`O-PR`) : This oracle crawler prioritizes the URLs according to PageRank scores.

3.5.4 Performance Metrics

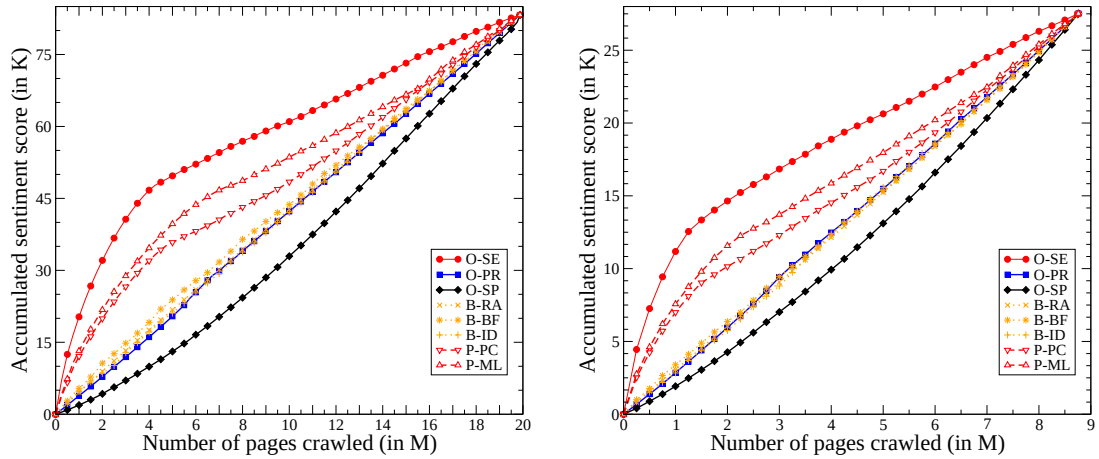
We simulate the crawlers with parameter combination `BP-WS-Adj`, which we conclude in Chapter 3.3 as the best one for ground-truth. The total sentimentality accumulated after fetching a certain number of pages is the main performance metric of our simulations, while PageRank values, spam scores, and the number of bytes downloaded are also reported for further analysis.

3.5.5 Seed Page Selection

In all simulations, we use 100 seed pages for the crawlers. According to the results reported in Table 3.9, using more seeds does not significantly increase the coverage of the crawler both for randomly selected seed pages and seed pages with highest outgoing links. We note that the lists for different seed types are totally disjoint, such that there is no page appearing in both lists.

3.6 Experimental Results

This section will present the results of experiments on sentimentality and polarity using random seeds and seed pages with highest outgoing links. We note that we have



(a) Sentiment accumulation without spam filtering (b) Sentiment accumulation with spam filtering

Figure 3.9: Sentimentality accumulation while pages are crawled from random seeds.

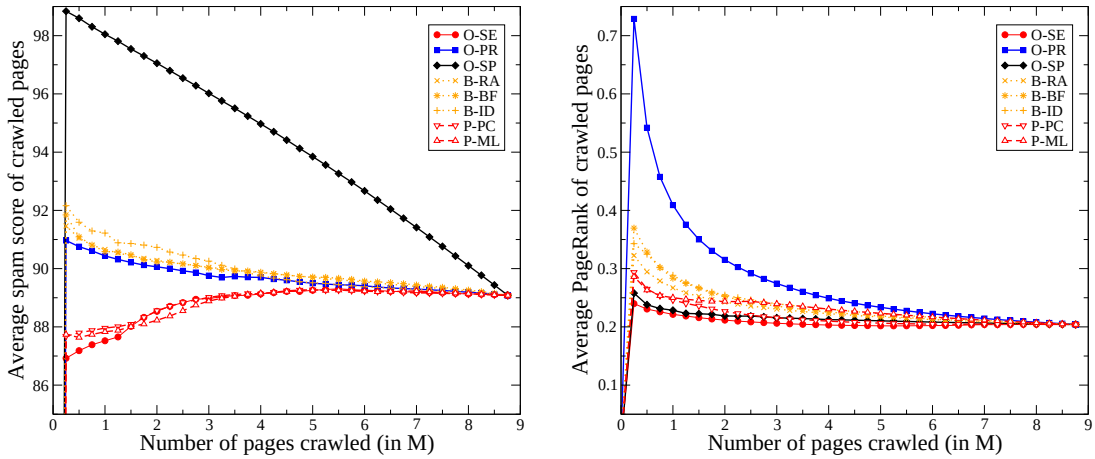
reported the results of experiments on sentimentality with a relatively smaller dataset in [87]. A detailed discussion on this experiment is given in Appendix A.

We also note that we do not report the results for sentiment prediction technique based on referring anchor text, as it performs similar to baseline crawlers, although it works fairly good in a small random dataset [87].

3.6.1 Experiments on Sentimentality

We first perform experiments using randomly selected seed pages. The observations of the total amount of sentimentality accumulated during the crawling process are shown in Figure 3.9, where spam filtering is off and on in sub-figures respectively. When spam filtering is on, any fetched page assumed as spam is ignored by the crawler. As previously shown in in Table 3.9, starting from the selected random seeds the crawler can access about 8.7M and 19.8M pages when spam filter is on and off respectively.

According to Figure 3.9(a), as expected, the oracle crawler (`O-SE`) achieves the best performance. This crawler can accumulate about two-third of the sentimentality available in the web page collection after crawling only less than half of the accessible pages. This finding justifies the second assumption we made in Section 3.1, i.e., the links in sentimental pages are likely to lead to other sentimental pages. All the proposed sentiment-focused crawlers (`P-PC, P-ML`) perform well approaching to performance of the oracle crawler (`O-SE`). In general, the remaining crawling strategies show relatively inferior performance in accumulating sentimentality.



(a) Average spam score accumulation

(b) Average PageRank accumulation

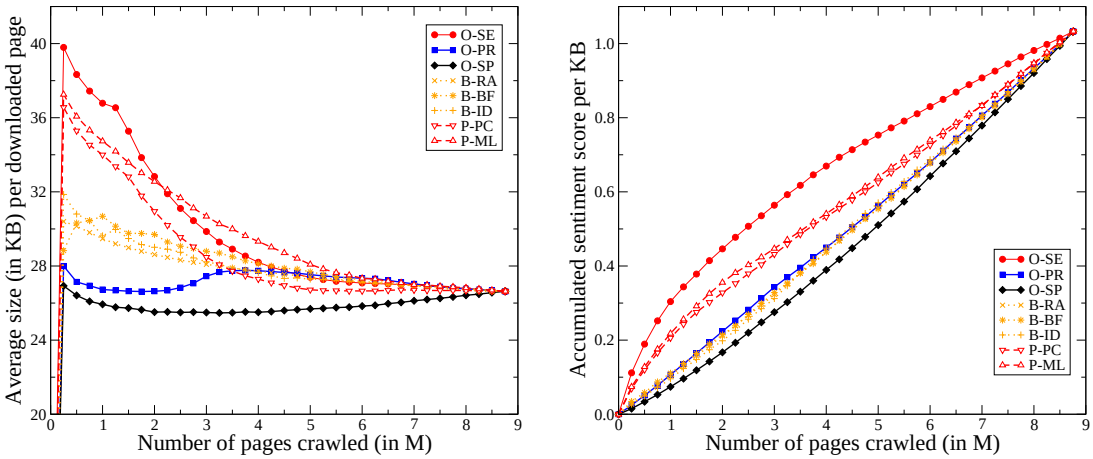
Figure 3.10: Spam score and PageRank accumulation without spam filtering while pages are crawled from random seeds.

As stated above, Figure 3.9(b) shows the total amount of sentimentality accumulated during the crawling process when the spam pages are ignored. Unsurprisingly, `O-SE` is the best performing strategy. `O-SE` accumulates almost half of the sentimentality available in the web page sample again after crawling only quarter of the accessible pages. Both of the proposed strategies outperform all baseline and oracle crawlers (other than `O-SE`) with a very good performance at early stages. As before, the `P-ML` strategy starts to perform well after the early stages of the crawling with respect to `P-PC`. All baselines (`B-BF`, `B-ID`, `B-RA`) perform better than `O-SP`, which prioritizes pages by their spam scores, but demonstrate an almost linear, steady performance.

Figure 3.10(a) displays the variation of the average spam score of the downloaded pages when the spam filter is present. As expected, the spam crawler `O-SP` quickly forms a collection with high spam scores. As seen, the sentiment-focused web crawling techniques do not create a negative bias since they do not lead to a significant increase in the spam rate of the crawled pages.

Figure 3.10(b) shows the average PageRank values of the downloaded pages when the spam filter is present. As expected, the PageRank oracle `O-PR` quickly forms a collection with high PageRank scores. The sentiment-focused crawling techniques also perform relatively good, which implies that the sentiment-focused web crawling techniques do not create a negative bias in terms of PageRank.

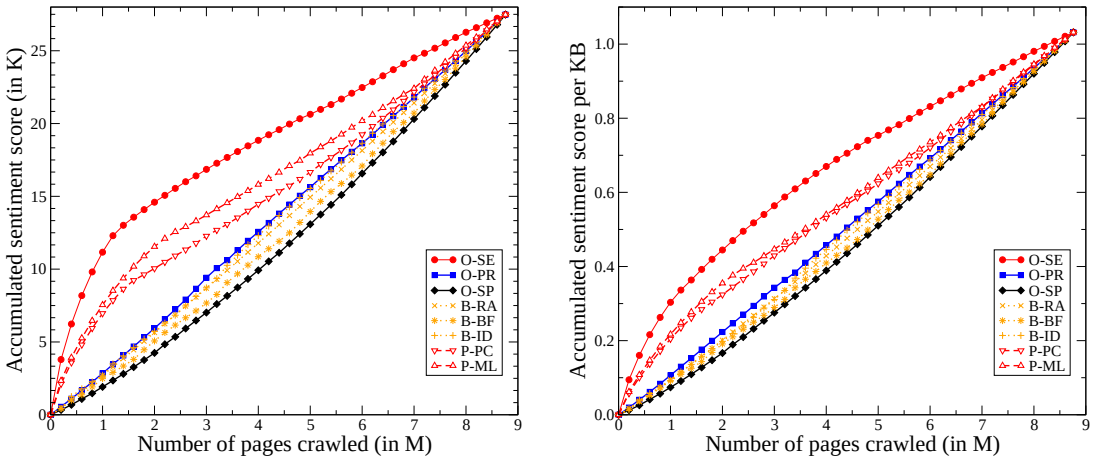
As the next performance metric, the variation of the crawled page sizes is displayed in Figure 3.11(a). According to the figure, `O-SE` and sentiment-focused web crawlers download slightly larger pages at the early stages compared to the other crawlers. It can be concluded that there is a correlation between the sentimentality score and the size of web pages in the sample data set.



(a) Average page size downloaded

(b) Sentiment per KB downloaded

Figure 3.11: Average page size and sentiment per KB downloaded without spam filtering while pages are crawled from random seeds.



(a) Sentiment accumulation

(b) Sentiment per KB downloaded

Figure 3.12: Results while pages are crawled with spam filtering from seeds with highest outgoing links.

As a final step for experiments with random seeds, we normalize the accumulated sentiment scores reported in Fig. 3.9(b) by the size of the pages (without removing HTML tags or anything else from the page). The normalized values are displayed in Figure 3.11(b), according to which there is no significant change in the performance gain achieved by the sentiment-focused web crawling strategies.

Before declaring the sentiment-focused crawling techniques as winner, we repeat the experiments using seed pages with highest outgoing links. Figure 3.12(a) shows the total amount of sentimentality accumulated during the crawling with spam filtering. The results are consistent with the observations using random seeds shown in Figure 3.9(b). As expected, the oracle crawler (`O-SE`) achieves the best performance. All the proposed sentiment-focused crawlers (`P-PC, P-ML`) perform well approaching to performance of

the oracle crawler (O-SE). When we observe the normalized sentiment score by the total size of the crawled pages in Figure 3.12(b), we state that the performance of crawling strategies consistent with sentiment accumulation in Figure 3.12(a).

The results of the experiments show that *i)* the links in sentimental pages are likely to lead to other sentimental pages, *ii)* the sentimentality of a page can be accurately estimated to a certain degree without having its content, and *iii)* there is no significant correlation between the PageRank and the sentimentality of a web page.

3.6.2 Experiments on Polarity

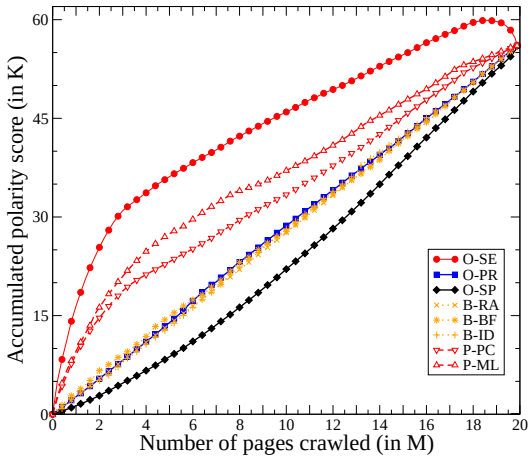
We further extend our experiments to investigate the applicability of the proposed ideas to polarity-focused web crawling. In this case, our objective is to discover pages including more positive content, as early as possible. In practice, early discovery of this kind of positive web content may be valuable for certain niche search engines, such as those specifically serving to children.

If we formulate the polarity-focused web crawling problem in accordance with the formulation in Section 3.1, our objective is to find a sequence $\langle p_{j_1}, \ldots, p_{j_i}, \ldots, p_{j_n} \rangle$ of n pages such that the total polarity $Pol(\mathcal{C}_n)$ of the pages in $\mathcal{C}_n$ is maximized, where each page $p_j \in \mathcal{C}_i$ is associated with a polarity score pol_j, i.e., we would like to maximize

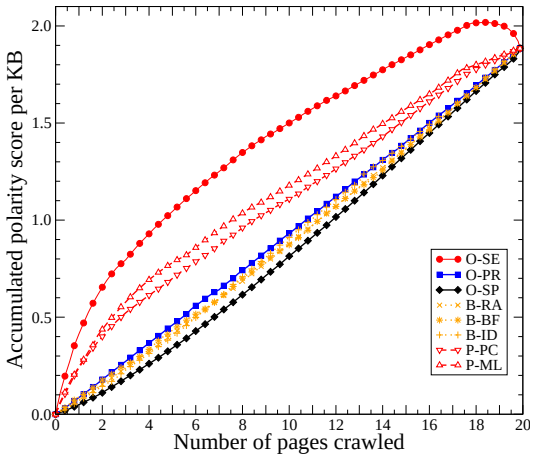
$$Pol(\mathcal{C}_n) = \sum_{i=1}^{n} pol_{j_i}. \tag{3.10}$$

In Figure 3.13, we report the polarity score accumulated together with the normalized polarity score by the total size of the crawled pages when the crawling is started with random seeds. In addition, Figure 3.14 reports the same information for crawling process with spam filtering using seeds with highest outgoing links. All these figures show similar trends. According to this set of figures, as expected, the oracle crawler that prioritizes pages by their actual polarity scores (O-PO) achieves the best performance. We observe a sharp decrease in the accumulated polarity scores towards the end of the crawling process. This is simply because the oracle crawler delays the download of pages with highly negative polarity scores till the very end of the process. However other crawlers, including the proposed ones, do not have a decreasing path, which indicates that most of the pages with negative polarity are fetched before the final stages.

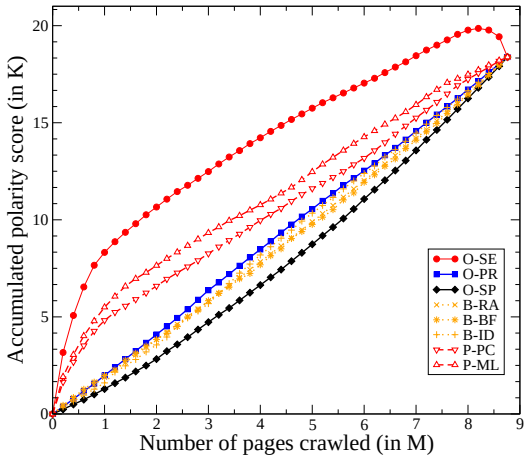
Similar to what we observed before in case of sentiment-focused crawling, the polarity-based crawling techniques (P-PC and P-ML) perform quite well with respect to the general-purpose crawling baselines. In summary, we can safely claim that polarity-focused web crawling could be a feasible technique in practice.



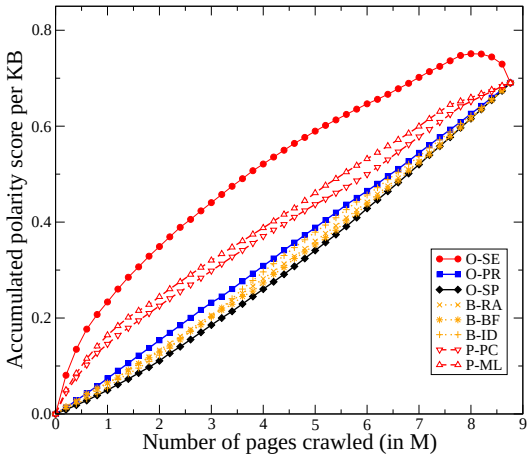
(a) Polarity accumulation without spam filtering



(b) Polarity per KB downloaded without spam filtering

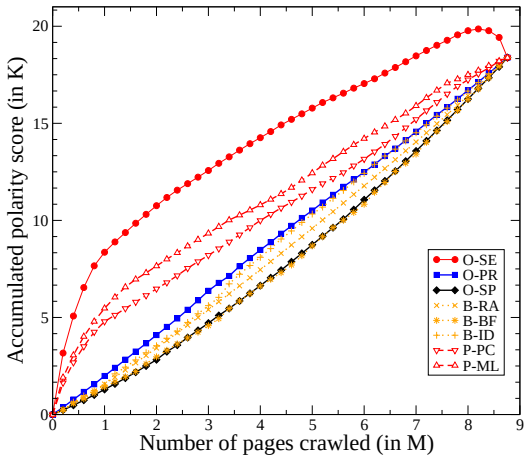


(c) Polarity accumulation without spam filtering

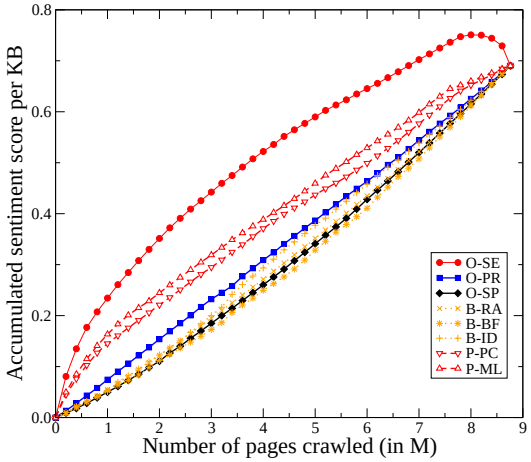


(d) Polarity per KB downloaded with spam filtering

Figure 3.13: Polarity results while pages are crawled from random seeds.



(a) Polarity accumulation



(b) Polarity per KB downloaded

Figure 3.14: Polarity accumulation while pages are crawled with spam filtering from seeds with highest outgoing links.

3.7 Discussion

In this chapter of the thesis, we introduced the sentiment-focused web crawling problem and designed a sentiment-focused web crawler for faster discovery and retrieval of sentimental context on the Web. In addition, we proposed different strategies for predicting the sentimentality and polarity scores of an "unseen" web page. We also conducted a user study to create a ground-truth for the sentimentality of web pages in our dataset. Finally, we compared the performance of our proposed strategies against baseline and oracle focused crawlers through simulations.

Our results for sentimental-focused web crawling are remarkable in the following sense. We have empirically shown that sentimentality scores of an English web page can be predicted to a certain degree without having the actual content of the page. The performance of proposed sentiment-focused web crawlers is shown to be better than general-purpose focused web crawlers in terms of early discovery of sentimental content, occasionally approaching the performance of the theoretical sentiment-focused crawler. In the light of these results, it can be concluded that sentiment-focused web crawling is feasible.

On the contrary, although anchor texts for URLs are good summaries on the referred pages in most of the cases, and Anh et al. [11] report that the use of anchor text in addition to the content significantly improves the effectiveness for ClueWeb09 retrieval, we experience that sentiment prediction technique based on referring anchor text, which works fairly good in a relatively small random dataset [87], does not scale for larger datasets. This may imply that most of the anchor texts do not include a sentiment word even if the referred page is sentimental. We plan to evaluate whether the performance of this prediction technique will improve when we take additional features (e.g. sentiment scores for N-Grams of target URL terms) into consideration.

Additionally, the proposed design and framework for sentiment-focused web crawler is extended for early discovery of web content with positive (or negative) polarity. Experiments have shown that the performance of our proposed sentiment-focused web crawler is better than general-purpose focused web crawlers.

As stated in Section 2.4, the only prior work that exploits the sentiment information in focused web crawling is the graph-based sentiment crawler by Fu et al. [42]. Our sentiment-focused web crawling approach differs from the graph-based sentiment crawler in certain ways.

- In [42], the sentiment information is combined with topical relevance to aid the retrieval of sentimental pages about a given target topic. In our work, the crawler is not focused to a particular topic. Instead, the goal is the early discovery and retrieval of (any kind of) sentimental content.

- Our work and [42] also differ in terms of design and evaluation of the proposed crawling techniques. The techniques in [42] are supervised and require retrieving/labeling large amounts of training data for every target topic. Our focused crawling framework does not require any training data or human effort (excluding the user study we conducted to tune the parameters of the crawler). Consequently, unlike the static classifiers in [42], our predictive model evolves in time as more pages are crawled and new models are learned.

- Because of the different nature of our problem, our evaluation is in terms of the total sentimentality gain and certain content quality metrics (e.g. spam and PageRank), instead of the relevance metrics used in [42] (e.g. precision and recall).

- The experiments of [42] involve intentionally selected set of relevant seed pages. On the contrary, we perform experiments using different seed selection techniques and for each seed type we empirically show that our proposed crawlers outperform baseline crawlers.

As a future work, we plan to extend our framework to be focused to a particular topic or a named-entity. Another potential future research direction is to identify new features that can further improve the sentiment prediction performance. It may also be interesting to extend our framework for other contexts, for instance a demographics-focused web crawler which aims to maximize the amount of demographics-relevant content in the crawled collection at early stages of crawling. Finally, we believe that our sentiment-focused web crawler framework has a high potential for constructing a commercial sentimental search engine since existing search engines (e.g. We Feel Fine) are sentiment specific such that they are incapable of presenting all kinds of sentimental content.

CHAPTER 4

SENTIMENT ANALYSIS IN TURKISH

"Kind[2] *words can be short and easy to speak, but their echoes are truly*[2] *endless."* [sentence: 2,-1] [result: max + and – of any sentence]

– Mother Teresa

While sentiment analysis has been an active research area for quite some time, unfortunately, most of the works are specific to the English language. Although there are recent commercial initiatives for emotional analysis of Turkish social media [3, 5], there exists limited number of academic works on sentiment analysis in Turkish. Being a highly agglutinative language makes sentiment detection and evaluation a complicated problem for Turkish.

This chapter presents a framework for unsupervised sentiment analysis in Turkish text documents. As part of the framework, using a state-of-the-art sentiment analysis library, SentiStrength is customized by translating its lexicon to Turkish. In addition, Zemberek [8] is utilized for morphological analysis of texts. The framework is applied to the problem of classifying the polarity of texts including movie reviews, hotel reviews and political news, obtained from popular Turkish social media sites. Although the framework is unsupervised, it is demonstrated to achieve fairly good classification accuracies, approaching the performance of supervised polarity classification techniques. Note that a preliminary version of this work appeared in [88].

The rest of this chapter is organized as follows. Section 4.1 introduces the sentiment analysis framework for Turkish. In Section 4.2, we evaluate our framework with experiments on polarity prediction of movie reviews, hotel reviews and political news. Section 4.3 discusses the caveats of our framework. Conclusive remarks are given in Section 4.4.

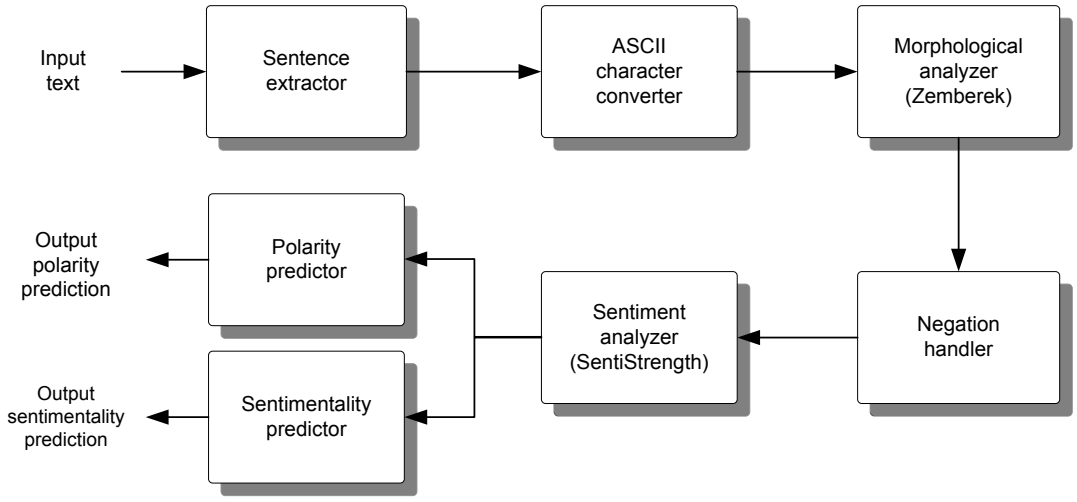


Figure 4.1: The pipeline of modules in the sentiment analysis framework.

4.1 Sentiment Analysis Framework

The motivation behind creating a sentiment analysis framework specific to Turkish, rather than using an existing framework for English, is due to certain differences between Turkish and English. These differences can be summarized as follows. First, Turkish is an agglutinative language, i.e., new and arbitrarily long words can be created by adding many suffixes to a root word. The added suffixes may change the polarity of words. In practice, it is not feasible to detect and add all variants of Turkish words into the sentimental word list. Second, negation words usually occur after the negated word. This is different than English, where negation words typically precede the word they negate. Moreover, in Turkish, the negation word can be in the form of a suffix ("-ma") within the word. Finally, Turkish has several letters that are missing in English ("ç", "ğ", "ı", "ö", "ş", "ü"). In informal writing on the Web, people tend to substitute these Turkish letters with the closest ASCII English letters ("c", "g", "i", "o", "s", "u"). This creates complication in identifying the words.

We have designed and implemented a sentiment analysis framework which takes into account the above-mentioned differences. Proposed framework consists of a pipeline of several software modules, each providing some input to the succeeding module in the pipeline. The input to the framework is a piece of text written in Turkish and the output is a prediction about the polarity and the sentimentality of the sentiments in the text. The pipeline for the proposed framework is illustrated in Figure 4.1. In what follows, we describe the modules in this pipeline.

- *Sentence extractor:* This is a simple module which splits the input text into sentences based on certain sentence separators (i.e., ".!?"). Each sentence is then passed to the next module as a separate input.

- *ASCII character converter:* Each word in the input sentence is looked up in a dictionary and checked for spelling errors. If a corresponding term is not found in the dictionary or there is a spelling error, the term is passed as input to an ASCII-tolerant parser to see if the word is written using ASCII character substitution. At this step, the parser may rewrite the term by substituting certain characters (e.g., "guzel" becomes "güzel").

- *Morphological analyzer:* Next, morphological analysis is performed on the words in the sentence. To this end, we use the Zemberek library, which is an open source, platform-independent, and general-purpose natural language processing library for Turkic languages.[1] Zemberek's morphological analyzer basically finds all possible root forms and suffixes of a given word. We always assume that the first morphological analysis result of Zemberek is the correct one and use that. After the morphological analysis, certain suffixes are removed from the selected word form. This is because some suffixes (e.g., tense and person suffixes) are not valuable for sentiment analysis.

- *Negation handler:* The negation takes places in Turkish most often in two forms, either in the form of a separate word negating one of the preceding words (e.g., "güzel değil" ("not nice")) or in the form of the "-ma" suffix, which is a part of the negated word (e.g., "olmayacak" ("it will not happen")). To handle the negations of the first form, we rely on a SentiStrength feature, which we will briefly describe later. To handle the second form of negations, we modify the sentence and introduce an artificial keyword ("_NOT_") before the negated word. This artificial word is added to the negation word list of our customized version of the SentiStrength library.

- *Sentiment analyzer:* As mentioned before, we customized the lexicon files of the SentiStrength library by translating them to Turkish. The translation is performed by human editors, who also added a few number of new words to the lists that were missing in the original SentiStrength. Table 4.1 shows the number of entries in the original (English) and customized (Turkish) versions of SentiStrength. Table 4.2 lists sample entries for sentiment word list in Turkish. Other than the changes in the lexicon files, we did not perform any modification in the scoring logic of SentiStrength as the codes of the library are not publicly available. To cope with the first form of negation words in Turkish, SentiStrength is initialized with a special parameter (`-negatingWordsOccurAfterSentiment`) to negate sentimental words before as well as after the negation words. SentiStrength by default applies negation to the words within a window of length

[1] Zemberek 2, `http://code.google.com/p/zemberek/`.

Table4.1: Number of lexicon entries in different lists of the original (English) and modified (Turkish) SentiStrength library

List	English version	Turkish version
Sentimental word list	2,546	1,366
Booster word list	27	13
Negation word list	16	4
Idiom list	9	39

Table4.2: Sample entries for SentiStrength's sentiment word list for Turkish

Word	Score	Word	Score
arzu	+4	abartı	−2
çekici	+2	aldırışsız*	−4
muhteşem	+4	diktatör*	−2
sevinç*	+3	eziyet	−5
süpriz	+1	tabu	−2
uyum	+2	tereddüt	−1

one. Our experiments indicated that a window size of three gives the best accuracy for Turkish, e.g., in the sentence "güzel film değil" ("it is not a nice movie"), "güzel" is affected from negation although it is not right before the negation word "değil" ("not").

- *Polarity predictor:* This module takes the sentiment scores associated with each word in the initial input text as well as the information about sentence splitting as input. The polarity of the input piece of text is determined according to the sentiment score assigned to the text. In this work, we evaluate three different approaches, which we refer to as `sen-bin`, `sen-max/min`, and `word-sum`:
  - `sen-bin`: For each sentence in the original input text, we use the binary score (i.e., +1 or −1) generated by SentiStrength. The sum of the scores over all sentences gives the sentiment score of the text.
  - `sen-max/min`: For each sentence, we use the maximum of positive word scores and the minimum of negative word scores that are computed by SentiStrength. We simply compute the average of these scores, separately for the positive and negative scores. The sum of the average of positive scores and the average of the negative scores give the sentiment score of the text.
  - `word-sum`: We use the sum of the sentiment scores of all words in the input text as its sentiment score.[2]

[2] The sentiment scores of individual word in the text can be obtained from SentiStrength via the `-explain` option.

Table4.3: The execution of the modules in the pipeline for a sample input text

Module	Output of the module
Original input text	"bu film cok guzel degildi. :(hic kimseye tavsiye etmem."
Sentence splitter	"bu film cok guzel degildi." and ":(hic kimseye tavsiye etmem."
ASCII converter	"bu film çok güzel değildi." and ":(hiç kimseye tavsiye etmem."
Morphological analyzer	"bu film çok güzel değil." and ":(hiç kimse tavsiye et-me-m."
Negation handler	"bu film çok güzel değil." and ":(hiç kimse tavsiye _NOT_ etmek."
Sentiment analyzer	"bu film çok güzel[3] değil [*-0.5 negated multiplier]." and ":([-1 emoticon]hiç kimse tavsiye[4] _NOT_ [*-0.5 negated multiplier] etmek."
Polarity predictor	(in all three methods, the polarity is predicted as negative)
(`sen-bin`)	-1 and -1
(`sen-max/min`)	$(+1, -2)$ and $(+1, -2)$
(`word-sum`)	-1.5 and -3
Sentimentality predictor	
(`abs-sen-max/min`)	1.0 $(+1, -2)$ and $(+1, -2)$
(`abs-word-sum`)	0.5 (4.5 with 9 words)

In all three scoring techniques, if the sentiment score of the text is positive, then its polarity is predicted as "positive"; otherwise, it is predicted as "negative". In this framework, we do not consider the neutral class and break the ties in favor of the negative class.

- *Sentimentality predictor:* This module is similar to the polarity predictor module. It takes as input the sentiment scores associated with each word in the initial input text as well as the information about sentence splitting. We evaluate two different approaches, which we refer to as `abs-sen-max/min` and `abs-word-sum`:
  - `abs-sen-max/min`: This is the sentimentality metric formulated by Kucuktunc et al. in [60]. For each sentence, we use the maximum of positive word scores and the minimum of negative word scores that are computed by SentiStrength. We simply compute the average of these scores, separately for the positive and negative scores. The scaled absolute sum of the average of positive scores and the average of the negative scores gives the sentimentality score of the text.
  - `abs-word-sum`: We first compute the absolute sum of the sentiment scores of all words in the input text, then use the average of this sum per words as its sentimentality score.

Table 4.3 shows the execution of these modules for a sample input text.

Table4.4: Properties of the movie review dataset used in the experiments

Property	Positive reviews	Negative reviews
Number of reviews	30,000	30,000
Reviews including an emoticon	4,093	2,477
Average number of words	36.02	37.07
Average number of sentences	3.75	3.82
Average word length	6.03	5.92

4.2 Experiments

To evaluate the sentiment analysis framework, we perform experiments on polarity prediction of online movie reviews (Section 4.2.1), hotel reviews (Section 4.2.2) and political news (Section 4.2.3) written in Turkish. A set of positive and negative examples for each dataset is given in the Appendix B.

In all experiments, the performance of proposed framework is evaluated in terms of accuracy, i.e., the ratio of the number of reviews whose polarity is correctly predicted to the total number of reviews. For the performance comparison, the baseline is defined as the evaluation of Google-translated[3] reviews with original SentiStrength tool, similar to approaches proposed by Bautin et al. [19] and Wan [89]. We note that Google Translate has the capability of "ASCII character conversion" for Turkish texts.

4.2.1 Polarity Detection of Movie Reviews

4.2.1.1 Dataset

The movie review data is obtained from a site called Beyazperde,[4] a well-known website that provides information about movies in Turkish. Beyazperde allows its users to comment on movies and state their opinion about the movie by selecting an icon (positive or negative), which forms the ground-truth polarity labels in our data. A sample screen of movie reviews from Beyazperde is shown in Figure 4.2. For the experiments, a random sample of positive and negative reviews, each with equal number of documents are picked. The statistics of the dataset are shown in Table 4.4.

4.2.1.2 Results

In our experiment, we activate all modules in the processing pipeline. Table 4.5 reports the accuracy values together with the true/false positive/negative rates for the three

[3] Google Translate, `http://translate.google.com/`
[4] Beyazperde, `http://www.beyazperde.com`.

Figure 4.2: Sample movie reviews from Beyazperde [38].

scoring techniques presented in Section 3.2. According to the table, all proposed scoring techniques perform considerably better than baseline. The `word-sum` scoring technique is the best performing technique, while `sen-bin` performs considerably worse than the other two scoring techniques. This result indicates that a fine-grain (at the word level) aggregation of the sentiment scores is more promising. In Table 4.5, we also observe that the prediction performance is better for the positive reviews. This is in contrast to what is reported by Thelwall et al. [84] for an English dataset. Overall, our performance does not reach the performance of Erogul's supervised machine learning approach on the same data (85% accuracy) [38]. However, given that our technique is unsupervised and independent of the problem domain, we believe that the accuracy achieved by our framework (78.5% accuracy) is promising.

Table 4.6 shows the accuracies when the ASCII conversion or morphological analysis modules are turned off. We note that turning off the morphological analysis also turns off the negation handling for the within-word negations. According to the table, most of the achieved accuracy is due to the customized SentiStrength library. Nevertheless, including the ASCII conversion and morphological analysis modules in the framework brings reasonable improvement, as seen in Table 4.6. In particular, for negative reviews, the accuracy increases by about 4% for all three scoring techniques when morphological analysis is turned on. Turning the ASCII conversion module on seems to help more in case of positive reviews.

Table4.5: Performance results (over all movie review instances)

Scheme	Metric	Baseline	Proposed
sen-bin	Accuracy	66.11%	71.08%
	True positive rate	31.35%	35.55%
	False positive rate	18.65%	14.45%
	True negative rate	34.76%	35.52%
	False negative rate	15.24%	14.48%
sen-max/min	Accuracy	70.63%	75.39%
	True positive rate	33.49%	38.41%
	False positive rate	16.52%	11.59%
	True negative rate	37.14%	36.98%
	False negative rate	12.86%	13.02%
word-sum	Accuracy	70.98%	76.49%
	True positive rate	34.66%	39.24%
	False positive rate	15.34%	10.76%
	True negative rate	36.32%	37.25%
	False negative rate	13.68%	12.75%

Table4.6: Accuracy when some modules are turned off for movie reviews

		Schemes		
Inst.	Modules	sen-bin	sen-max/min	word-sum
All	All modules	71.08%	75.39%	76.49%
	No ASCII conversion	70.41%	74.66%	75.71%
	No morphological analysis	68.60%	72.49%	73.51%
+	All modules	71.05%	76.82%	78.49%
	No ASCII conversion	70.06%	75.19%	76.89%
	No morphological analysis	70.02%	75.54%	76.86%
-	All modules	71.05%	73.96%	74.50%
	No ASCII conversion	70.76%	74.12%	74.53%
	No morphological analysis	67.19%	69.45%	70.17%

We note that the performance results reported in this section are better than the results given in our previous study [88]. Although the same framework is used for the experiments, the customized lexicon file for SentiStrength has been improved to include 1,366 entries, about 500 entries more than the one used in [88].

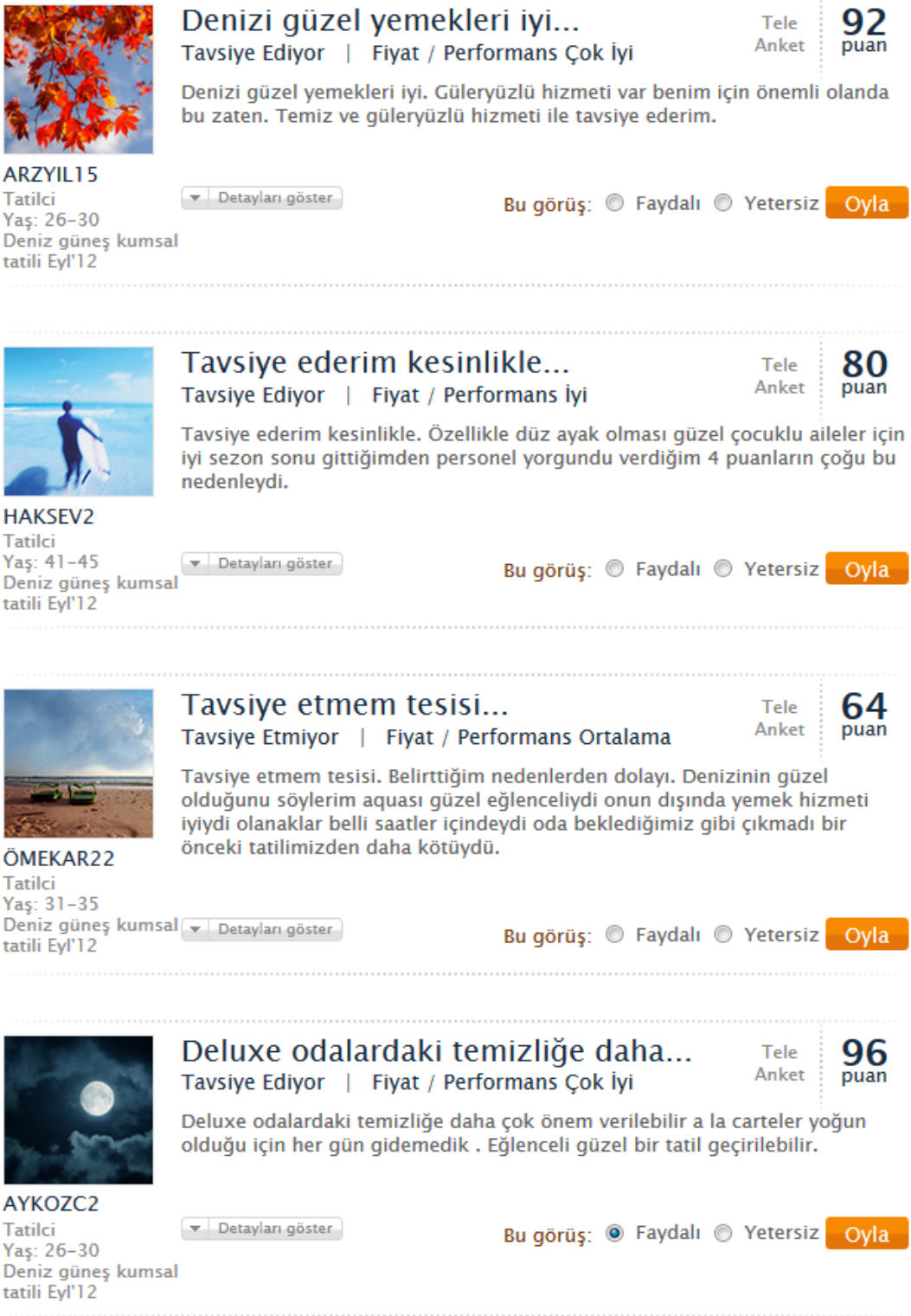


Figure 4.3: Sample hotel reviews from OtelPuan.

Table4.7: Properties of the hotel review dataset used in the experiments

Property	Positive reviews	Negative reviews
Number of reviews	70,627	15,605
Reviews including an emoticon	412	94
Average number of words	23.14	24.20
Average number of sentences	3.42	3.42
Average word length	6.13	6.05

4.2.2 Polarity Detection of Hotel Reviews

4.2.2.1 Dataset

The hotel reviews data is obtained from a hotel review site called OtelPuan [5], a popular website that provides information about more than 2,000 hotels in Turkey. The dataset belongs to years 2006 to 2012.

A sample screen of hotel reviews from OtelPuan is shown in Figure 4.3. Registered users are able to criticize and evaluate the facilities by writing reviews and giving scores between 0-100. In addition, they state their opinion on whether they recommend the hotel as "Tavsiye Ediyor" (positive), "Tavsiye Etmiyor" (negative) or "Tavsiye Kararsız" (neutral), regardless of the overall score given. The properties of the dataset used in experiments are given in Table 4.7.

4.2.2.2 Results

Table 4.8 reports the accuracy values for the three scoring techniques mentioned in Section 3.2, while all modules in the processing pipeline are activated. According to the table, all proposed scoring techniques perform considerably better than baseline except for the negative reviews. The `word-sum` scoring technique is again the best performing technique. This result supports the inference of Section 4.2.1.2 that a fine-grain (at the word level) aggregation of the sentiment scores is more promising. In addition, we again observe that the prediction performance is considerably better for positive reviews. Overall, our performance (83.8% accuracy) is promising.

Table 4.9 shows the accuracies when the ASCII conversion or morphological analysis modules are turned off. According to the table, most of the achieved accuracy is again due to the customized SentiStrength library. For negative reviews, the effect of using morphological analysis is noticeably high. This indicates that negative reviews mostly include word negations. Another interesting finding for negative reviews is that turning off for ASCII conversion improves the accuracy, which may indicate that there are some conflicting terms used in negative terms.

[5] OtelPuan, `http://www.otelpuan.com`.

Table4.8: Performance results (over all hotel review instances)

Scheme	Metric	Baseline	Proposed
sen-bin	Accuracy	70.86%	76.98%
	True positive rate	36.24%	40.11%
	False positive rate	13.76%	9.89%
	True negative rate	31.76%	31.14%
	False negative rate	18.24%	18.86%
sen-max/min	Accuracy	77.00%	82.77%
	True positive rate	39.27%	43.40%
	False positive rate	10.73%	6.60%
	True negative rate	35.00%	32.26%
	False negative rate	15.00%	17.74%
word-sum	Accuracy	76.47%	83.76%
	True positive rate	38.95%	43.69%
	False positive rate	11.05%	6.31%
	True negative rate	35.02%	33.68%
	False negative rate	14.98%	16.32%

Table4.9: Accuracy when some modules are turned off for hotel reviews

		Schemes		
Inst.	Modules	sen-bin	sen-max/min	word-sum
All	All modules	76.98%	82.77%	83.76%
	No ASCII conversion	75.54%	80.81%	81.67%
	No morphological analysis	70.64%	75.83%	76.65%
+	All modules	80.22%	86.81%	87.38%
	No ASCII conversion	78.23%	83.93%	84.43%
	No morphological analysis	76.46%	83.69%	84.42%
-	All modules	62.28%	64.51%	67.37%
	No ASCII conversion	63.40%	66.72%	69.20%
	No morphological analysis	44.31%	40.27%	41.49%

4.2.3 Polarity Detection of Political News

4.2.3.1 Dataset

This is the political news dataset defined by Kaya et al. in [56]. The dataset is collected from 6 different Turkish newspapers and annotated as positive or negative by three native speakers of Turkish. There are 200 positive and 200 negative texts in the dataset, of which all three annotators are agreed on polarity. Kaya et al. stated that annotating (therefore analysing) political news is a challenging task. First, the polarity of a sentence in a political news depends highly on the political view of the annotator. Second, most of the time columnists put positive and negative criticisms together.

Table4.10: Properties of the political news dataset used in the experiments

Property	Positive reviews	Negative reviews
Number of reviews	200	200
Reviews including an emoticon	6	4
Average number of words	462.74	474.38
Average number of sentences	43.92	49.03
Average word length	6.32	6.17

The properties of the dataset used in experiments are given in Table 4.10. As seen, the texts of this dataset are much longer than the previous review datasets.

4.2.3.2 Results

Kaya et al. in [56] compared four supervised machine learning algorithms of Naïve Bayes, Maximum Entropy, SVM and the character based N-Gram Language Model using this dataset. Using the claim of [73] which states that to classify the texts it might suffice to produce a list of sentimental words that people tend to use to express strong sentiments, Kaya et al. created a baseline using 197 positive and 300 negative indicators, reporting 59% accuracy. In experiments, all machine algorithms clearly surpass the baseline and it is reported that Maximum Entropy (when using the frequency of features) and N-Gram Language Model (when using the presence of features) perform better than other algorithms in terms of accuracy at 76-77%.

Table 4.11 reports the accuracy values for the three scoring techniques mentioned in Section 3.2, while all modules in the processing pipeline are activated. In terms of accuracy, all our techniques perform better than both the Google translated baseline and the baseline given in in [56], but poorer than the performance of supervised methods reported by Kaya et al. in [56]. Interestingly, `sen-max/min` scoring technique is the best performing one among our proposed techniques. Although `word-sum` is better than `sen-max/min` for negative news, it performs worst for positive news. Kaya et. al reported that most of the time columnists put positive and negative criticisms together. This may indicate that word level aggregation of the sentiment scores fails for positive news as they include a substantial amount of words with negative sentimentality.

Comparing the precision and recall values reported in [56], it can be stated that supervised techniques of Kaya et al. have a better accuracy for positive news. On the contrary, for all our techniques, the prediction performance is better for negative news, which is the case reported by Thelwall et al. [84] for an English dataset.

One of the findings of Kaya et al. [56] is that the positive effect of adjectives and effective words on the sentiment classification in news domain cannot be generalized.

Table4.11: Performance results (over all political news instances)

Scheme	Metric	Baseline	Proposed
sen-bin	Accuracy	62.00%	67.50%
	True positive rate	35.50%	56.00%
	False positive rate	64.50%	44.00%
	True negative rate	88.50%	79.00%
	False negative rate	11.50%	21.00%
sen-max/min	Accuracy	65.50%	71.25%
	True positive rate	37.50%	60.50%
	False positive rate	62.50%	39.50%
	True negative rate	93.50%	82.00%
	False negative rate	6.50%	18.00%
word-sum	Accuracy	64.00%	70.50%
	True positive rate	33.50%	55.50%
	False positive rate	66.50%	44.50%
	True negative rate	94.50%	85.50%
	False negative rate	5.50%	14.50%

To verify this finding, we perform the experiments using lexicon file including only sentimental adjectives. The average accuracy value as 61% justifies that adjectives do not carry a great deal of information for the sentiments of news.

To sum up, we believe that the accuracy achieved by our unsupervised framework (71.25% accuracy) is again promising.

4.3 Caveats

Our framework is not perfect. This section summarizes some of the issues encountered when using the linguistic tools in our framework.

- Since the reviews (especially the movie reviews) are written using rather an informal language, the likelihood of typos is high. Most often the typos are intentional and are hard to detect by linguistic tools. The spell-checking algorithm we used can handle up to three misplaced or wrong characters in the root and two misplaced or wrong characters in the word suffixes. This check improves the accuracy up to only 0.3%, but the execution time of the overall system increases about 10 times.
- As stated before, some of the reviews are written using ASCII letters, replacing original Turkish letters. The ASCII converter we used cannot always detect the correct version of the words, as it has no domain knowledge. For instance, "yas" ("mourning") having a negative sentiment score of -3 is the ASCII version of "yaş" ("age") which has neutral sentiment. Our conversion tool fails in certain cases like this.

- We observed difficulties in sentence splitting due to the misuse of punctuation. This results in wrong interpretations of the negation. For example, the review "filmde hic bir sey yok. manasiz bir film", which contains negative sentiments, gains a positive sentiment if the periods are omitted as in "filmde hic bir sey yok manasiz bir film" ("yok" negates "manasiz").

- Our framework does not perform named-entity recognition. So when a person name (like "Yiğit" or "Sevgi") or a film name having sentimental words is referred in a review, our framework cannot exclude such named-entity from sentimental analysis.

- As stated before, our technique is unsupervised and independent of the problem domain. However, especially the hotel reviews include domain-specific negative sentimental words that our framework does not detect. Some sample sentimental words for hotel review domain are "uzun kuyruk" ("long queue"), "çakıl sahil" ("shore with gravel"), "ücret(li)" ("fee"). In addition, in our lexicon there are negative sentiments, i.e. "korku" ("horror"), "çığlık" ("scream") in the lexicon which would indicate a positive sentiment in movie reviews (e.g., a good horror movie should be scary). This context-dependency is a problem for polarity prediction, but does not affect sentimentality prediction unless the sentiment turns into neutral.

4.4 Discussion

In this chapter, we proposed a framework for unsupervised sentiment analysis in Turkish text documents. Our framework uses various linguistic tools as well as the customized version of the SentiStrength sentiment analysis tool. We evaluated the performance of our framework by applying it to the problem of polarity prediction of movie reviews, hotel reviews and political news. The experiments over different social media datasets with relatively large corpus indicate reasonable prediction accuracy. We observe that word level aggregation of the sentiment scores is working well especially for short texts whereas sentence level aggregation is more promising for polarity prediction of longer texts.

The customized SentiStrength library for Turkish is freely available to the research community at the official website of SentiStrength.

As the future work, first we plan to improve our framework by providing solutions to issues listed in Section 4.3. We also plan to evaluate the performance of our framework over other social media datasets (such as Twitter). In addition, to optimize the sentiment word strengths, the framework will be tested against human evaluated texts.

CHAPTER 5

TURKISH SENTIMENT-FOCUSED WEB CRAWLING

"Ne mutlu[2] *Türküm diyene."* [sentence: 2,-1] [result: max + and – of any sentence]

– Mustafa Kemal Atatürk

The proposed framework in Chapter 3 can only work with web pages in English, since the sentiment analysis tool used in the implementation supports only English. This chapter introduces a modified version of sentiment-focused web crawler capable of crawling Turkish web pages, facilitating the sentiment analysis framework for Turkish proposed in Chapter 4. In Section 5.1, we introduce the modified framework for Turkish support. Section 5.2 presents the results of the user study conducted in order to create a ground-truth for sentimentality in Turkish web pages. We provide details on dataset characteristics and experimental setup in Section 5.3. The experimental results are given in Section 5.4. Finally, we discuss our findings and point to future work in Section 5.5.

5.1 Framework

As Turkish sentiment focused crawler is a modified version of sentiment-focused web crawler proposed in Chapter 3, it facilitates and improves the high-level architecture presented in Figure 3.1. The architecture of Turkish sentiment focused crawler is composed of the same main components of the generic architecture:

- **Page retrieval component**: This component fetches the web pages from the Web by picking up the URLs from the download queue and detects the language of the web page.
- **Storage component**: This component is responsible for storing the retrieved web pages in the Content Database if the detected language of the page is Turkish.

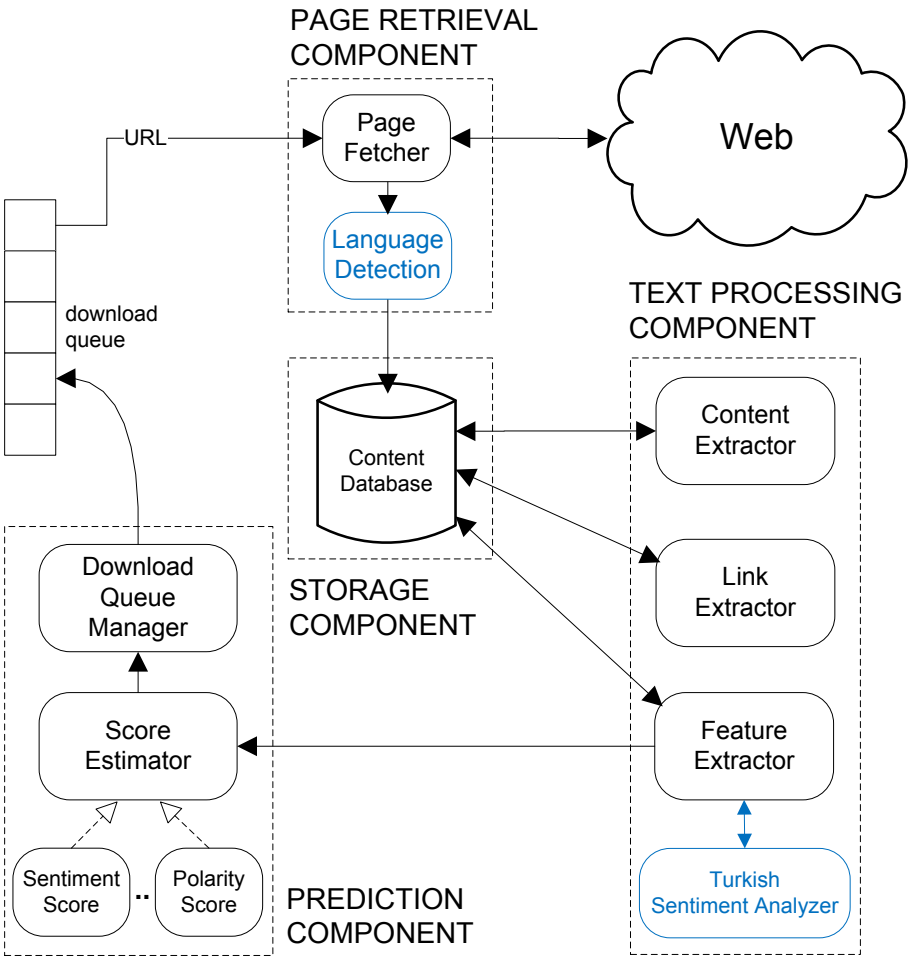


Figure 5.1: Generic architecture for Turkish sentiment-focused crawling.

- **Text processing component**: This component processes each retrieved web page to extract features. First, textual content is extracted by parsing and cleaning. The obtained textual content is then tokenized into sentences and words. Second, the URLs that are linked by the page are extracted. Finally, features associated with the textual content, URLs, and anchor text, are extracted. For the features related to sentimentality, our Turkish sentiment analysis framework is used.

- **Prediction component**: This component is composed of two sub-modules: score estimator and download queue manager. The main task of score estimator is to predict the sentimentality (or polarity) for extracted URLs using the features provided by the feature extractor. After assignment of sentiment (or polarity) scores, extracted links are given to the download queue manager, which reorganizes the URLs in the download queue in decreasing order of the predicted scores. The URL with the highest sentiment (or polarity) score is passed to the page fetcher as the next URL to be downloaded.

Table5.1: Number of lexicon entries in different lists of the original (English) and modified (Turkish) SentiStrength library

	English version		Turkish version	
List	Positive	Negative	Positive	Negative
Sentimental word list	397	2,249	432	934
Sentimental adjective list	145	299	174	412

Table5.2: Parameter combinations for Turkish sentiment-focused crawling framework

	Text Extraction	Sentiment Score	Lexicon (Turkish version)
HP-SS-All$_{TR}$	HTML Parser	sentence scores	all words
HP-SS-Adj$_{TR}$	HTML Parser	sentence scores	sentimental adjectives
HP-WS-All$_{TR}$	HTML Parser	word scores	all words
HP-WS-Adj$_{TR}$	HTML Parser	word scores	sentimental adjectives
BP-SS-All$_{TR}$	BoilerPipe	sentence scores	all words
BP-SS-Adj$_{TR}$	BoilerPipe	sentence scores	sentimental adjectives
BP-WS-All$_{TR}$	BoilerPipe	word scores	all words
BP-WS-Adj$_{TR}$	BoilerPipe	word scores	sentimental adjectives

The architecture of Turkish sentiment focused crawler illustrated in Figure 5.1. As explained previously, text processing component of the architecture includes the Turkish sentiment analysis framework (see Section 4.1 for details) for extraction of sentimentality related features for the web page. The statistics of lexicon file used by the sentiment analysis framework are shown in Table 5.1.

5.2 User Study

As stated in Section 3.3, we need to know the actual sentimentality scores of the fetched pages in order to evaluate the performance of different focused crawling techniques. However, we do not have a ground-truth for sentiment scores of a Turkish web page collection. Consequently, we again choose to generate some score estimates that substitute the actual sentiment scores.

Since we modify the generic architecture for sentiment focused web crawling, we have the same couple of alternatives for tools in the framework and methods for sentimentality prediction as described in Section 3.3, with the exception that lexicon entries are in Turkish. Therefore, this yields to again eight possible parameter combinations to be considered as summarized in Table 5.2.

Table5.3: The degree of agreement among the judges in terms of overlapping

		Overlap (O)							
Judge	S	J1	J2	J3	J4	J5	GT1	GT2	GT3
J1	0.33	1.00	0.78	0.82	0.82	0.83	0.78	0.90	0.81
J2	0.27	0.78	1.00	0.83	0.78	0.75	0.73	0.87	0.87
J3	0.26	0.82	0.83	1.00	0.77	0.75	0.72	0.86	0.87
J4	0.39	0.82	0.78	0.77	1.00	0.83	0.84	0.90	0.75
J5	0.49	0.83	0.75	0.75	0.83	1.00	0.95	0.85	0.65
Avg.	0.35	0.85	0.83	0.83	0.84	0.83	0.80	0.87	0.79

Table5.4: The degree of agreement among the judges in terms of Kappa values

		Kappa (κ)							
Judge	S	J1	J2	J3	J4	J5	GT1	GT2	GT3
J1	0.33	1.00	0.47	0.46	0.60	0.66	0.58	0.77	0.49
J2	0.27	0.47	1.00	0.56	0.51	0.49	0.48	0.69	0.60
J3	0.26	0.46	0.56	1.00	0.49	0.49	0.46	0.66	0.61
J4	0.39	0.60	0.51	0.49	1.00	0.67	0.69	0.78	0.40
J5	0.49	0.66	0.49	0.49	0.67	1.00	0.89	0.69	0.28
Avg.	0.35	0.64	0.60	0.60	0.65	0.66	0.62	0.72	0.48

A small-scale user study with 5 judges (J1, J2, J3, J4, and J5) is conducted to identify the best performing parameter combination listed in in Table 3.1. For the user study, a sample of 500 web pages is randomly chosen from tour Turkish web page collection (see Section 5.3.1 for details). The judges individually evaluate and assign the labels "sentimental" or "not sentimental" to a page. Based on the labeling of the judges, same three ground-truths defined in Section 3.3 are used:

- **Ground Truth-1** (GT1): In this scenario, a page is assumed to be sentimental if at least one judge thinks so.
- **Ground Truth-2** (GT2): This scenario works according to majority voting. A page is assumed to be sentimental when the majority of judges think so.
- **Ground Truth-3** (GT3): This scenario is the most strict one. A page is labeled as sentimental only if all the judges agree on the sentimentality of the page.

Table 5.3 and Table 5.4 display the agreement between judges as well as their agreement with the ground-truths. About one-third of the pages are labelled as sentimental on average. According to the table, the overlap (O) between judges is above 83 %, which signs a high agreement between the judges. The kappa values (κ) also support this

Table5.5: The ranking quality achieved by different parameter combinations over 500 randomly sampled pages in Turkish

GT	Metric	Rand	HP-SS-All$_{TR}$	HP-WS-All$_{TR}$	BP-SS-All$_{TR}$	BP-WS-All$_{TR}$	HP-SS-Adj$_{TR}$	HP-WS-Adj$_{TR}$	BP-SS-Adj$_{TR}$	BP-WS-Adj$_{TR}$
	P@10	0.54	0.80	0.80	0.80	0.70	0.60	**1.00**	**1.00**	**1.00**
	P@50	0.54	0.76	0.90	0.84	0.88	0.78	0.86	0.90	**0.96**
GT1	P@100	0.54	0.82	0.84	0.86	0.86	0.81	0.87	0.90	**0.93**
	AP	0.54	0.76	**0.78**	0.77	0.77	0.73	0.77	0.75	0.77
	DCG	38.24	41.28	41.11	40.97	41.46	40.86	41.83	41.76	**41.99**
	P@10	0.34	0.50	0.40	0.60	0.30	0.30	0.50	0.70	**0.80**
	P@50	0.34	0.62	0.68	0.66	0.60	0.48	0.62	0.64	**0.70**
GT2	P@100	0.34	0.67	0.65	0.67	0.61	0.58	0.66	0.69	**0.71**
	AP	0.35	0.59	0.57	0.58	0.55	0.51	0.55	0.58	**0.59**
	DCG	24.41	27.38	26.53	26.99	26.21	25.80	26.57	**27.79**	27.25
	P@10	0.13	0.20	0.10	0.30	0.20	0.10	0.30	**0.50**	0.40
	P@50	0.13	0.18	**0.38**	0.28	0.20	0.16	**0.38**	0.30	0.30
GT3	P@100	0.13	0.28	0.34	0.31	0.24	0.25	0.30	0.32	**0.35**
	AP	0.14	0.24	0.29	0.28	0.25	0.23	0.28	**0.30**	**0.30**
	DCG	9.59	10.43	10.93	10.97	10.60	10.20	11.10	**11.51**	11.35

Table5.6: The ranking quality observed for individual parameter alternatives

GT	Metric	HP	BP	SS	WS	All$_{TR}$	Adj$_{TR}$
	P@10	0.80	0.88	0.80	0.88	0.78	0.90
	P@50	0.83	0.90	0.82	0.90	0.85	0.88
GT1	P@100	0.84	0.89	0.85	0.88	0.85	0.88
	AP	0.76	0.76	0.75	0.77	0.77	0.75
	DCG	41.27	41.54	41.22	41.59	41.20	41.61
	P@10	0.43	0.60	0.53	0.50	0.45	0.58
	P@50	0.60	0.65	0.60	0.65	0.64	0.61
GT2	P@100	0.64	0.67	0.65	0.66	0.65	0.66
	AP	0.55	0.57	0.56	0.56	0.57	0.55
	DCG	26.57	27.06	26.99	26.64	26.78	26.85
	P@10	0.18	0.35	0.28	0.25	0.20	0.33
	P@50	0.28	0.27	0.23	0.32	0.26	0.29
GT3	P@100	0.29	0.31	0.29	0.31	0.29	0.31
	AP	0.26	0.28	0.26	0.28	0.26	0.27
	DCG	10.66	11.11	10.78	11.00	10.73	11.14

inference. As a result of the high inter-judge agreement, it can be concluded that the labels obtained through the user study form a sufficiently reliable basis to evaluate the performance of the parameter combinations.

Table5.7: Sample Turkish seed pages for different types of categories

Category	Web page	URL	Domain
auto	Auto Show	http://www.autoshow.com.tr	com
blog	Cartalete	http://cartalete.blogspot.com	com
computer	Pardus	http://www.pardus.org.tr	org
education	ODTÜ	http://www.metu.edu.tr	edu
finance	İş Bankası	http://www.isbank.com.tr	com
food	Oktay Usta	http://www.oktayustam.com	com
fun	Zaytung	http://www.zaytung.com	com
government	TC Kültür Bakanlığı	http://www.kultur.gov.tr	gov
jobs	Kariyer.net	http://www.kariyer.net	net
health	Depresyon	http://www.depresyon.info.tr	info
media	Anadolu Ajansı	http://www.aa.com.tr	com
military	Kara Kuvvetleri Komutanlığı	http://www.kkk.tsk.tr	tsk
music	CSO	http://www.cso.gov.tr	gov
newspaper	Cumhuriyet Gazetesi	http://www.cumhuriyet.com.tr	com
people	Kim Kimdir	http://www.kimkimdir.gen.tr	gen
politics	CHP	http://www.chp.org.tr	org
religion	Diyanet İşleri Başkanlığı	http://www.diyanet.gov.tr	gov
science	Elektrik.gen.TR	http://www.elektrik.gen.tr	gen
shopping	HepsiBurada	http://www.hepsiburada.com	com
sports	NTV Spor	http://www.ntvspor.net	net
travel	Otel Puan	http://www.otelpuan.com	com
wiki	Wikipedia	http://tr.wikipedia.org	org

Table 5.5 provides the performance and correctness measures, each generated using a different parameter combination. In the table, values given as **bold** correspond to the best results for the related metric. As seen, all of the rankings perform considerably better than the baseline. The `BP-WS-Adj`$_{TR}$ combination yields the most accurate rankings for the relatively relaxed `GT1` and the majority `GT2` scenarios. Interestingly, `BP-SS-Adj`$_{TR}$ combination performs slightly better than `BP-WS-Adj`$_{TR}$ combination for the more conservative `GT2` scenario.

Table 5.6 summarizes the performance of individual parameter alternatives (with the same logic applied for Table 3.5). According to table, it is interesting to note that all three types of parameters again have a winner and all winners are same with the alternatives in original sentiment focused web crawling described in Section 3.3. BoilerPipe again performs slightly better than HTML Parser, which shows that it is better on HTML content cleaning even for Turkish web pages. The word-level sentiment scores again provide significant improvement rather than the sentence-level scores. Similar to the observations in Section 3.3, this result indicates that a fine-grain (at the word level) aggregation of the sentiment scores is more promising for evaluating sentimentality. Finally, the lexicon containing adjectives yields a better performance rather than using the translated original lexicon entries. It can be stated that the default SentiStrength lexicon has superfluous amount of sentimental words, some of which are context-dependent.

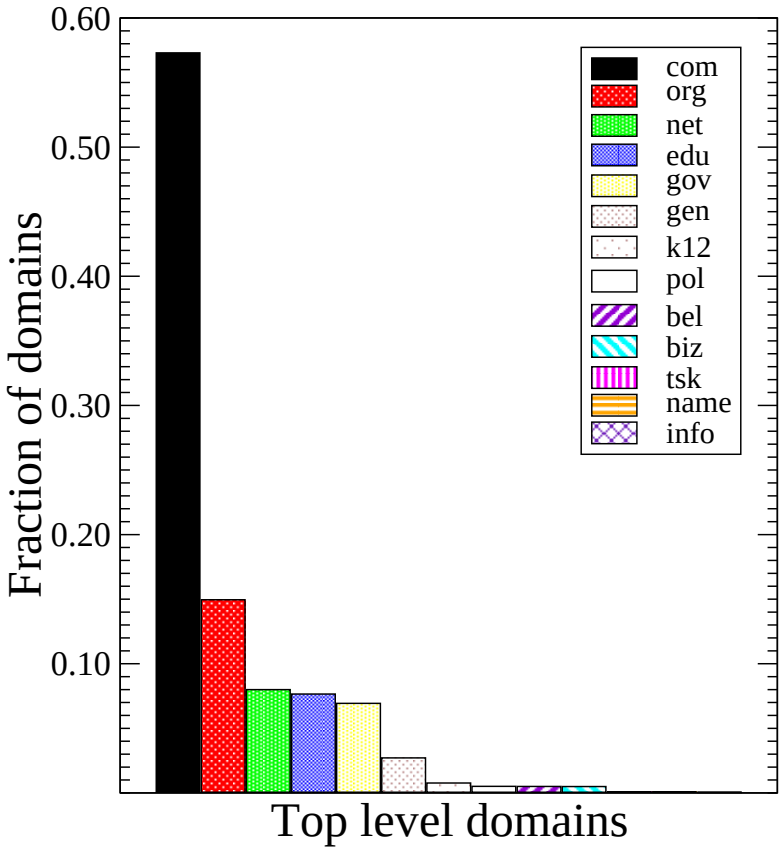


Figure 5.2: Fractions of domains in Turkish web collection.

As a result of all these findings, it can be concluded that BP-WS-Adj$_{TR}$ is best alternative as the ground-truth for our Turkish web page collection.

5.3 Experimental Setup

5.3.1 Dataset

Unfortunately, there is no publicly available web collection of Turkish pages, which entails us to create our own web collection. The web collection is crawled starting from more than 400 seed pages in Turkish, selected from Alexa[1], Open Directory Project[2], and top results to domain searches by Google Search Engine[3]. A list of sample seed pages for different types of categories are given in Table 5.7. The total number of pages for a site (including its sub-domains) is limited to 4,000.

[1] Alexa, http://www.alexa.com
[2] DMOZ, http://www.dmoz.org
[3] Google, http://www.google.com.tr

Table5.8: Sentimentality and polarity distribution statistics for experiment dataset

Distribution	Min.	Avg.	Max.	Std. dev.
Sentiment score	0.00	0.0003	0.8	0.0129
Polarity score	−0.75	0.0008	0.8	0.0124

In order to crawl web pages, we customized an open source Java crawler, Crawler4j [78] which provides a simple interface for crawling the Web while obeying the politeness policies. In order to detect the language of fetched web pages, a language detection library implemented in Java is used [79]. This language detection library detects language of a text using naive Bayesian filter and reports that it succeeds 99% over precision for 53 languages, including Turkish.

As result of crawling, we have a web collection of 789,349 web pages in Turkish. The average size of a page in our dataset is 63,326 and 3,461 bytes before and after removing HTML tags, respectively. A page contains 603.9 words (276.8 unique words) and 18.5 sentences, on average. About 9.4% of pages do not link to any other page, and about 25.4% of them do not receive a link from any other page. There are 147.1 outbound and 70.6 inbound links per page. Majority of the pages belong to `com` domain with rate 57.3%. The fractions of web pages per domains are given in Figure 5.2.

Comparing the statistics of Turkish web collection with ClueWeb09 collection given in Section 3.5.1, we have the following findings:

- While HTML size of Turkish pages are about twice size of English pages, the text sizes are nearly same.
- Since the total number of words in a Turkish page is less while there are more sentences in the text after HTML content cleaning, Turkish sentences are shorter than English sentences.
- The number of outbound and inbound links in Turkish pages are two times more than the number of links in English pages. This may be a result of news and forum pages which include excessive number of links to other inter-site pages.
- The fractions of top level domains are almost same while the majority of the pages belongs `com` domain.

We associate every page in the dataset with two values: sentiment score and polarity score. Unlike to ClueWeb09 dataset, we can not associate pages with spam score and PageRank since we do not have these scores for our Turkish web collection. Table 5.8 displays the minimum, average, and maximum values as well as the standard deviations observed for the sentiment and polarity score distributions in our dataset.

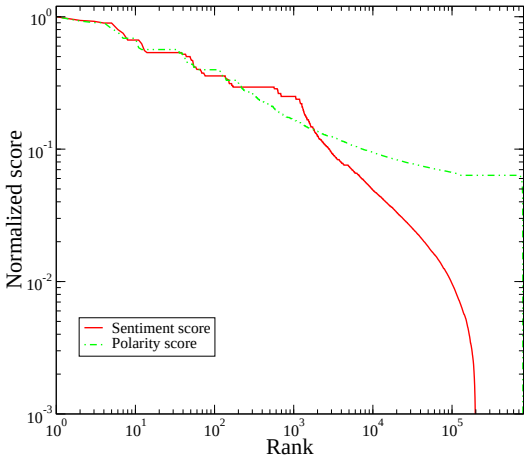


Figure 5.3: Normalized sentiment and polarity score distributions in Turkish web collection (log-log scale).

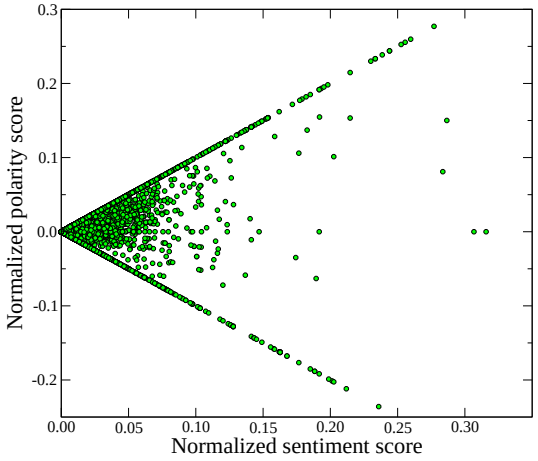


Figure 5.4: Sentiment score versus polarity score.

The distribution of the sentiment and polarity scores with respect to the rank of a page in the distribution is shown in Figure 5.3. In this figure, the ranges for each score type are normalized to [0,1] range. In terms of skewness, polarity score has the most skewed distribution. Figure 5.4 illustrates that there is no correlation between sentiment and polarity scores of a page.

5.3.2 Setup

For the experiments, the framework given in Section 5.1 is simulated with an implementation in Java 6.0. MySQL (version 5.1.61) database is used for storing the web pages and the extracted features. Experiments are executed on a 16-core computer with 48GB of RAM running Debian Linux. On this setup, the execution time required for experiments varies from 1 to 2 hours.

5.3.3 Performance Metrics

In all simulations, we set the seed page size to 100 pages for comparability of results with original sentiment-focused web crawler given in 3.6. In addition, increasing the size of set does not significantly affect the coverage of crawl as shown in Table 5.9 especially for seed pages with highest-outgoing links.

As the main performance metric, we compute the total sentimentality and polarity accumulated reported at regular intervals, after 1,000 pages are crawled. The data values in plots are down-sampled for better visibility.

Table5.9: Coverage of pages with different sizes and types of seed pages

Seed size	Random	Highest outgoing link
1	1,470	176,236
10	180,039	181,912
20	185,407	181,912
50	203,229	181,912
100	226,502	181,912
200	253,381	181,912

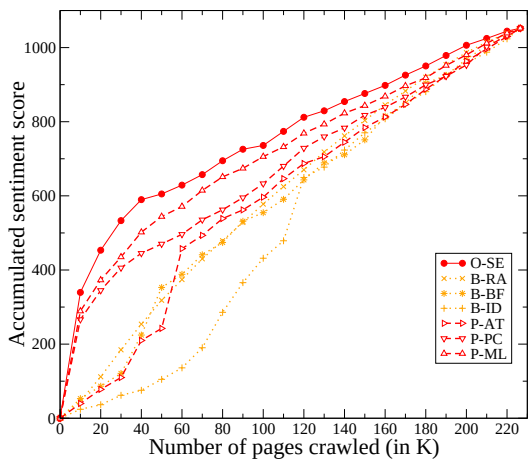
5.3.4 Crawlers

For Turkish sentiment focused web crawling, we do not define or propose new crawlers, but utilize the sentiment-focused web crawlers (`(P-AT)`, `(P-PC)`, `(P-ML)`) and baseline crawlers (`(B-RA)`, `(B-ID)`, `(B-BF)`) together with the theoretical sentiment focused crawler `(O-SE)` described in Section 3.5.3.

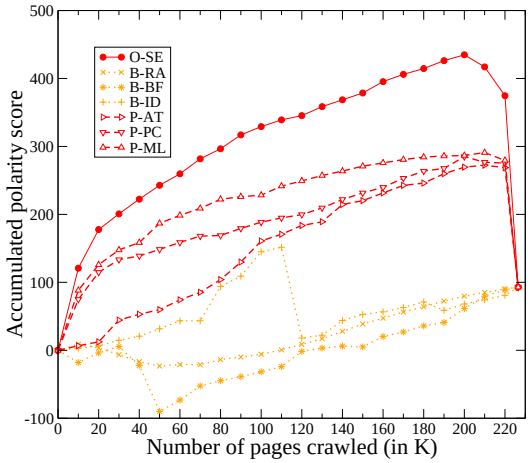
5.4 Experimental Results

We start our crawling simulations using random seed pages. The crawlers can access about 226K pages starting from the random seeds. Figure 5.5 shows the results of crawling for sentimentality (on the left side) and polarity (on the right side). According to Figure 5.5(a) and (b), as expected, the oracle crawlers `(O-SE)` that prioritize pages by their actual scores achieve the best performance, accumulating half of the sentimentality and polarity after crawling only one-fifth of the accessible pages. The proposed sentiment-focused crawling techniques also perform quite well both in sentiment and polarity accumulation. At the early stages of crawling, `P-PC` and `P-ML` perform similar while `P-AT` is poorer even compared to random crawler `B-RA`. At some point, the performance of `P-AT` becomes better than baseline crawlers but its overall performance for sentimentality accumulation is not promising. For polarity accumulation, all proposed techniques fetch most of the pages with negative polarity at the final stages as expected, whereas `B-BF` and `B-ID` accumulate such negative pages at early stages in bulk. The performance gap of `P-PC` and `P-ML` compared with the oracle is not very large for sentimentality accumulation, i.e., the predicted scores are good substitutes for the actual scores. On the contrary, there is still gap for improvement for proposed techniques in polarity accumulation.

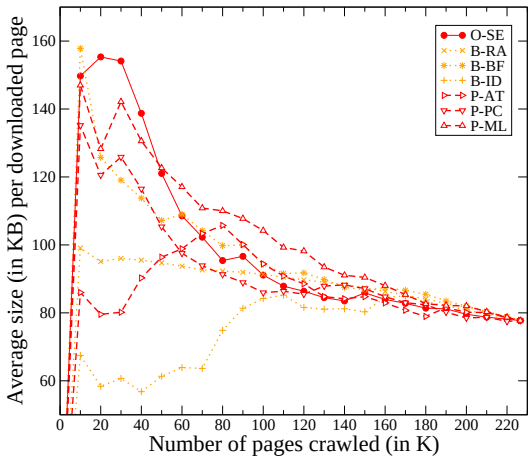
Figure 5.5(c) and (d) reports the variation results of the crawled page sizes. According to these figures, sentiment-focused web crawlers except `P-AT`, download slightly larger pages at the early stages compared to the other crawlers. Figure 5.5(e) and (f) display the normalized accumulated scores (reported in Figure 5.5(a) and (b) respectively)



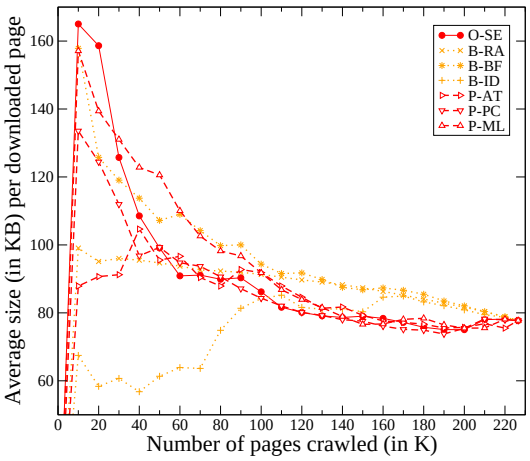
(a) Sentiment accumulation



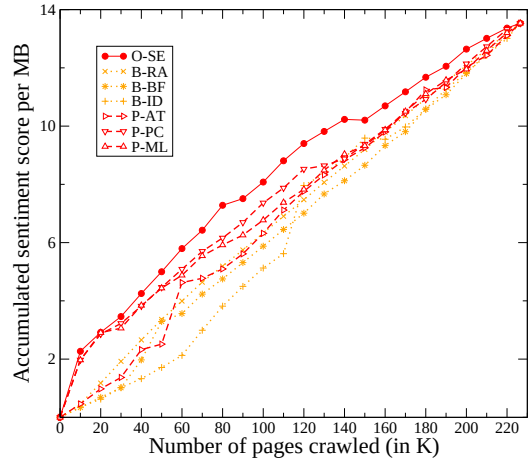
(b) Polarity accumulation



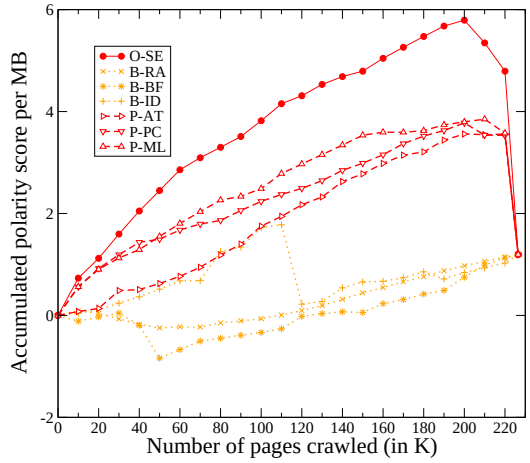
(c) Average page size for sentiment crawl



(d) Average page size for polarity crawl



(e) Sentiment per MB Download



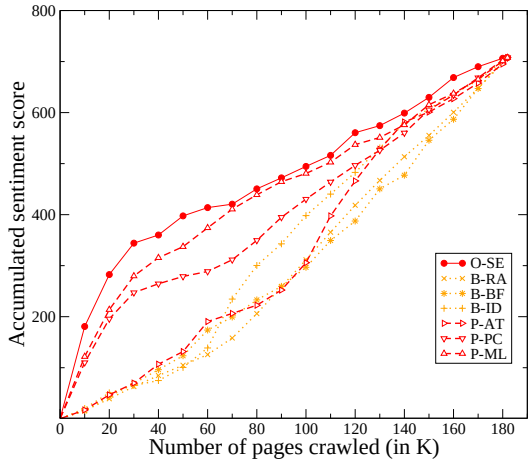
(f) Polarity per MB Download

Figure 5.5: Results while pages are crawled from random seeds.

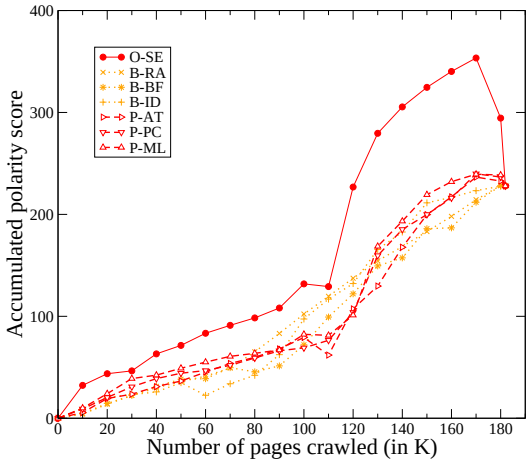
by the size of the pages. The results of normalized accumulated scores are consistent with the accumulation results in Figure 5.5(a) and (b), and from these results, we can conclude that although the sentiment-focused crawlers are good at reaching sentimental content early they are slower in doing so.

In the next step, we execute our crawling experiments using seed pages with highest outgoing links. There is no intersection between random seed pages and seed pages with highest outgoing links. The crawlers can access about 182K pages starting from the seeds with highest outgoing links. Figure 5.6 shows the results of crawling for sentimentality (on the left side) and polarity (on the right side). According to Figure 5.6(a) and (b), the oracle crawlers `(O-SE)` again achieve the best performance as expected although their performances are different. While `(O-SE)` accumulates half of the sentimentality after crawling only one-third of the accessible pages, its accumulation performance for polarity is poor as half of the polarity is accumulated after crawling more than two-third of the accessible pages. Moreover, the performances for proposed sentiment-focused crawling techniques differ for sentiment and polarity accumulation. For sentiment accumulation, the performances for `(P-ML)` and `(P-PC)` seem similar to results retrieved with random seeds whereas accumulation speeds are relatively low. On the contrary, they have an ordinary performance similar to baseline crawlers in polarity accumulation. The normalized the accumulated scores shown in Figure 5.6(e) and (f) verify this situation as well. These findings indicate that there is a problematic issue with the seed pages. When we examine the polarity scores of seed pages, the majority of the seed pages are neutral but the rest has a negative polarity. This implies that pages with negative polarity tend to link to neutral or negative pages, and the link structure between sentimental pages does not allow quick discovery positive pages starting from seed pages with negative polarity.

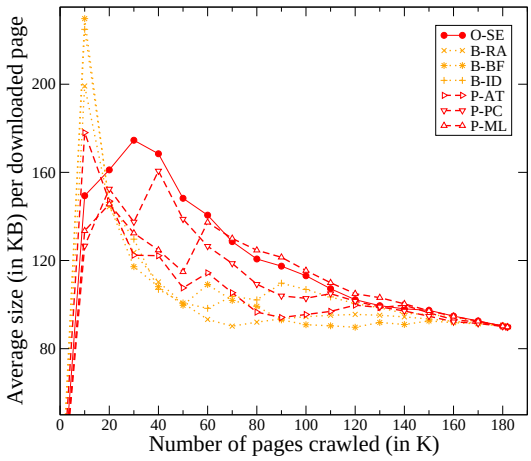
Finally, to investigate the effect of polarity scores of seeds on crawling for polarity, we execute our crawling experiments using seed pages with highest (positive) polarity scores. The crawlers can access about 198K pages starting from the seeds with highest outgoing links. Figure 5.7 shows the results of crawling for sentimentality (on the left side) and polarity (on the right side). According to Figure 5.7(a) and (b), the oracle crawlers `(O-SE)` again achieve the best performance, especially for polarity accumulation. The performances of proposed techniques are not promising for sentiment accumulation. This may indicate that polarity does not imply sentimentality and most of the sentimental pages in the dataset contain both positive and negative sentimental words. For polarity accumulation, `(P-ML)` outperforms other proposed and baseline crawlers. When we examine the performance of `(P-AT)`, it can be stated that the anchor text alone is not sufficient to predict the sentimentality and polarity of a target page. The results of normalized the accumulated scores (given in Figure 5.7(e) and (f)) are again consistent with the accumulation results in Figure 5.7(a) and (b).



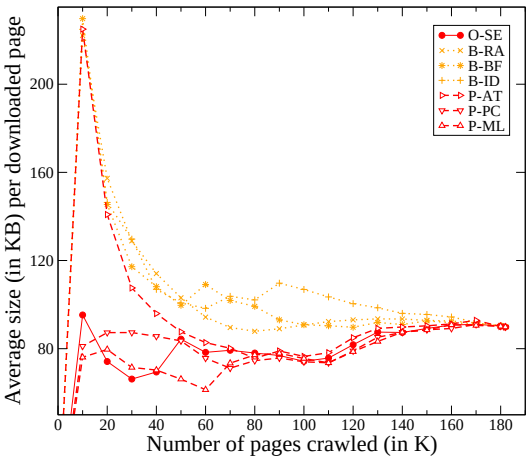
(a) Sentiment accumulation



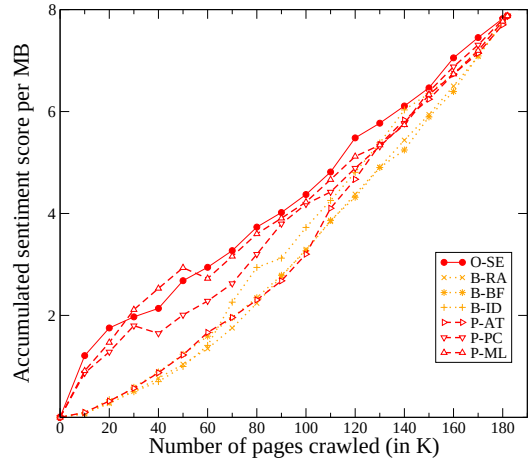
(b) Polarity accumulation



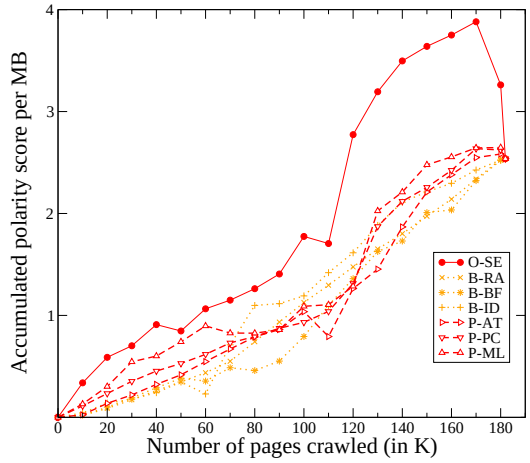
(c) Average page size for sentiment crawl



(d) Average page size for polarity crawl

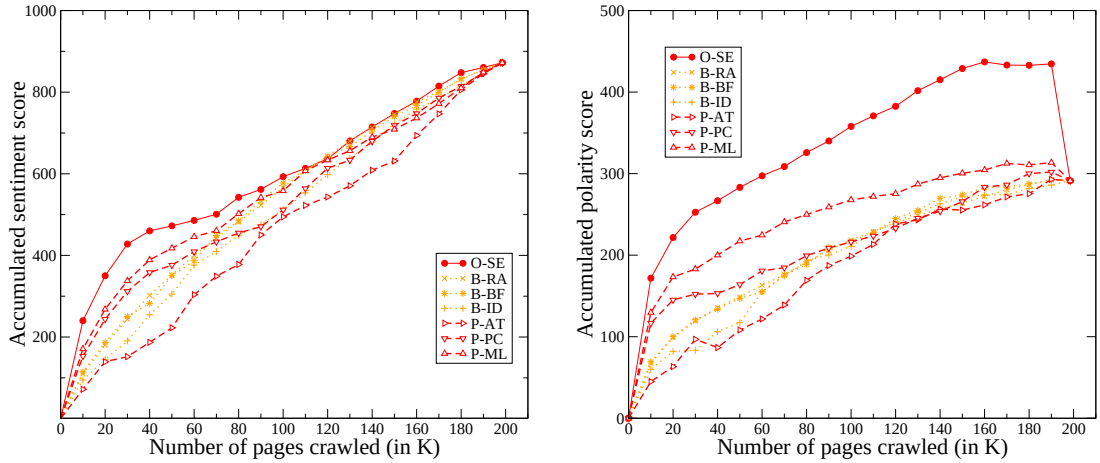


(e) Sentiment per MB Download



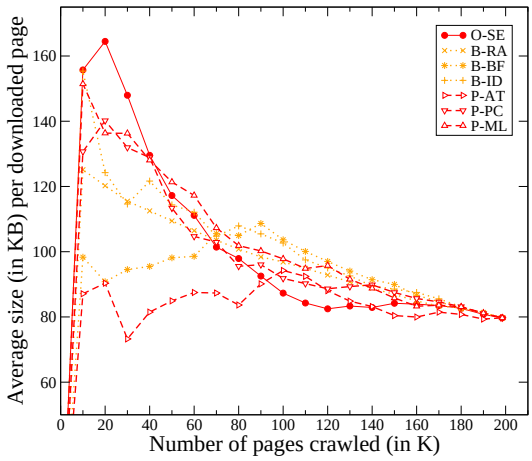
(f) Polarity per MB Download

Figure 5.6: Results while pages are crawled from seeds with highest outgoing links.

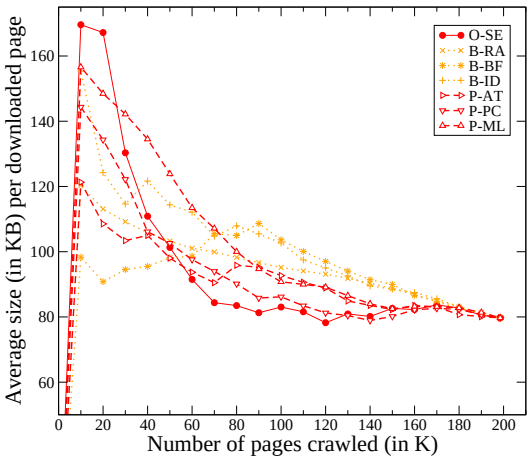


(a) Sentiment accumulation

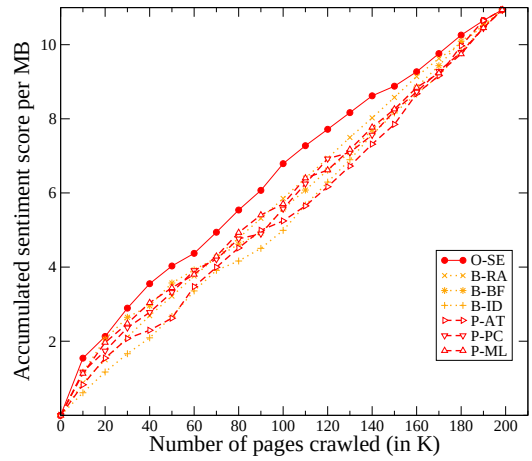
(b) Polarity accumulation



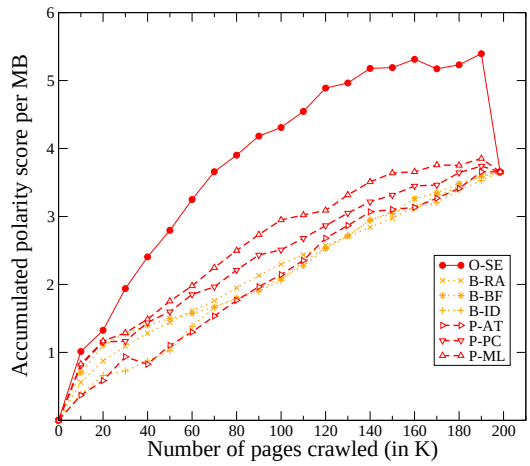
(c) Average page size for sentiment crawl



(d) Average page size for polarity crawl



(e) Sentiment per MB Download



(f) Polarity per MB Download

Figure 5.7: Results while pages are crawled from seeds with highest polarity scores.

5.5 Discussion

This chapter presented a framework for sentiment-focused web crawling for Turkish by modifying the framework and using the proposed strategies for sentimentality prediction given in Chapter 3. This framework is the first Turkish sentiment-focused web crawler design, using the state-of-the-art page processing and sentiment analysis tools. We believe the proposed work is beneficial in that it will enable efficient solution the quick discovery of sentimental content in Turkish. The main findings of this work are as follows.

- Sentiment-focused web crawling is feasible also for Turkish pages as the link structure between sentimental pages allow quick discovery of such pages.
- The amount of sentiments expressed in a Turkish web page can be predicted to a certain degree even before obtaining the content of the page.
- Sentiment-focused web crawling beats traditional crawling alternatives in terms of the early discovery and retrieval of sentimental content from the Web.
- As observed for sentiment-focused web crawling in Chapter 3, only the anchor texts on the links alone are not sufficient to predict the sentimentality of a target page.

The following are among our future development plans. We plan to increase the size of Turkish web collection and repeat the experiments. Moreover, we will make our dataset available for academic research in order to support research on information retrieval and related language technologies on Turkish.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

"I think and think for months and years, ninety-nine times, the conclusion is false. The hundredth time I am right."

– Albert Einstein

Recently, the analysis of textual content having sentiments and opinions towards entities, has gained importance due to its high potential for monetization. Despite the vast interest on sentiment analysis in terms of extraction, classification, summarization, and presentation, somewhat surprisingly, the discovery of the sentimental content is mostly ignored. In addition, most of the works on sentiment analysis are specific to English and there exists limited number of works for Turkish. This thesis aims to fill these gaps.

First, we introduced the sentiment-focused web crawling problem, and proposed a sentiment-focused web crawling framework for faster discovery and retrieval of sentiment/opinion sources on the Web. In addition, we proposed different strategies for predicting the sentimentality and polarity scores of an "unseen" web page. We also conducted a user study to create a ground-truth for the sentimentality of web pages in our dataset. We compared the performance of proposed strategies against baseline and oracle focused crawlers through simulations. We further extended the proposed design and framework for sentiment-focused web crawler for early discovery of web content with positive (or negative) polarity. We showed with empirical evidence that sentimentality and polarity scores of a web page can be predicted to a certain degree without having the actual content of the page. As the performance of proposed sentiment-focused web crawler is shown to be better than general-purpose focused web crawlers in terms of early discovery of sentimental content, occasionally approaching the performance of an oracle sentiment-focused crawler, we concluded that sentiment-focused web crawling is feasible.

Second, we proposed a sentiment analysis framework for Turkish using various linguistic tools as well as the customized version of the SentiStrength sentiment analysis tool. We achieved reasonable prediction accuracy on experiments over different social media datasets. We observed that word level aggregation of the sentiment scores is working well especially for short texts whereas sentence level aggregation is more promising for polarity prediction of longer texts.

Finally, we extended our sentiment-focused web crawling framework for Turkish by utilizing our sentiment analysis framework for Turkish. We observed that proposed sentiment-focused web crawler beats traditional crawling alternatives in terms of the early discovery and retrieval of sentimental content from the Web. With the empirical evidence we showed that sentiment-focused web crawling is feasible for Turkish web pages.

There are some future work directions regarding the contributions of this thesis. Sentiment-focused web crawling has a high potential for constructing a commercial sentimental search engine. In addition, we currently conduct studies on extending our sentiment-focused web crawling framework for demographics (e.g. gender, age) context, such that we aim to maximize the amount of demographics-relevant content in the crawled collection at early stages of crawling. The main problem is that we lack the ground-truth for demographics relevance scores of web pages. On the contrary, we believe that our idea is feasible since there are promising accuracy results reported on gender and age prediction using words in the pages visited [50]. For our Turkish sentiment analysis framework, we plan to provide solutions to our caveats and evaluate the framework over other social media data (e.g. Twitter). For our Turkish sentiment-focused web crawling framework, we will repeat the experiments with larger web collection. To do so, we plan to increase the size of our Turkish web collection and make it available for academic research.

REFERENCES

[1] Opinion Crawl - sentiment analysis tool for the web and social media, http://www.opinioncrawl.com, last visit on September 1, 2013.

[2] The size of the World Wide Web, http://www.worldwidewebsize.com, last visit on September 1, 2013.

[3] SkyKeeper, http://skykeeper.t2.com.tr, last visit on September 1, 2013.

[4] Stats - WordPress.com, http://en.wordpress.com/stats/, last visit on September 1, 2013.

[5] Türkçe otomatik sosyal medya takibi ve analizi, http://www.etkitakip.com, last visit on September 1, 2013.

[6] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26(3):12:1–12:34, 2008.

[7] M. Abdul-Mageed, M. T. Diab, and M. Korayem. Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 587–591, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] A. A. Akin and M. D. Akin. Zemberek, an open source NLP framework for Turkic languages. 2007.

[9] J. Alpert and N. Hajaj. We knew the web was big, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, last visit on September 1, 2013.

[10] I. S. Altingovde and O. Ulusoy. Exploiting interclass rules for focused crawling. *IEEE Intelligent Systems*, 19(6):66–73, Nov. 2004.

[11] V. N. Anh and A. Moffat. The role of anchor text in ClueWeb09 retrieval. In *TREC*, 2010.

[12] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, Aug. 2001.

[13] V. Atteveldt, J. Kleinnijenhuis, N. Ruigrok, and S. Schlobach. Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations. *Journal of Information Technology & Politics*, 5(1):73–94, 2008.

[14] A. E. S. Baccianella and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, pages 2200–2204, 2010.

[15] R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, and F. Silvestri. Challenges on distributed web retrieval. In *Proceedings of the 23rd International Conference on Data Engineering*, pages 6–20, 2007.

[16] X. Bai. Predicting consumer sentiments from online text. *Decision Support Systems*, 50(4):732–742, 2011.

[17] A. Balahur and M. Turchi. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 52–60, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[18] S. Batsakis, E. G. M. Petrakis, and E. Milios. Improving the performance of focused web crawlers. *Data & Knowledge Engineering*, 68(10):1001–1013, October 2009.

[19] 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*, 2008.

[20] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. Exploring sentiment summarization. In *Proc. AAAI Spring Symp. Exploring Attitude and Affect in Text: Theories and Applications*, pages 1–4, 2004.

[21] F. Benamara, S. Irit, C. Cesarano, N. Federico, and D. Reforgiato. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of International Conference on Weblogs and Social Media*, 2007.

[22] P. D. Bra, G.-J. Houben, Y. Kornatzky, and R. Post. Information retrieval in distributed hypertexts. In *Proceedings of the 4th International Conference on Computer-Assisted Information Retrieval*, pages 481–493, 1994.

[23] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International Conference on World Wide Web*, WWW '07, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[24] J. Brooke, M. Tofiloski, and M. Taboada. Cross-linguistic sentiment analysis: From English to Spanish. In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria*, pages 50–54, 2009.

[25] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, June 1996.

[26] S. Chakrabarti, K. Punera, and M. Subramanyam. Accelerated focused crawling through online relevance feedback. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 148–159, New York, NY, USA, 2002. ACM.

[27] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623–1640, 1999.

[28] C.-C. Chang and C.-J. Lin. LibSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[29] S. Chelaru, I. S. Altingovde, and S. Siersdorfer. Analyzing the polarity of opinionated queries. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, ECIR'12, pages 463–467, Berlin, Heidelberg, 2012. Springer-Verlag.

[30] J. Cho and H. Garcia-Molina. Estimating frequency of change. *ACM Transacations on Internet Technology*, 3(3):256–290, Aug. 2003.

[31] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1-7):161–172, Apr. 1998.

[32] Y. Choi, K. Kim, and M. Kang. A focused crawling for the web resource discovery using a modified proximal support vector machines. In *Proceedings of the International Conference on Computational Science and its Applications - Volume Part I*, pages 186–194, 2005.

[33] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.

[34] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pages 519–528, 2003.

[35] G. Demartini and S. Siersdorfer. Dear search engine: What's your opinion about...?: Sentiment analysis for semantic enrichment of web search results. In *Proceedings of the 3rd International Semantic Search Workshop*, SEMSEARCH '10, pages 4:1–4:7, New York, NY, USA, 2010. ACM.

[36] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. Focused crawling using context graphs. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 527–534, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[37] M. Ehrig and A. Maedche. Ontology-focused crawling of web documents. In *Proceedings of the ACM Symposium on Applied Computing*, pages 1174–1178, 2003.

[38] U. Erogul. Sentiment analysis in Turkish. Master's thesis, Middle East Technical University, 2009.

[39] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422, 2006.

[40] R. Feldman. Techniques and applications for sentiment analysis. *Communications of ACM*, 56(4):82–89, Apr. 2013.

[41] L. A. Freitas and R. Vieira. Ontology based feature level opinion mining for Portuguese reviews. In *Proceedings of the 22nd International Conference on World Wide Web companion*, WWW '13 Companion, pages 367–370, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[42] T. Fu, A. Abbasi, D. Zeng, and H. Chen. Sentimental spidering: Leveraging opinion information in focused crawlers. *ACM Transactions on Information Systems*, 30(4):24, 2012.

[43] M. Ganapathibhotla and B. Liu. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 241–248, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[44] S. Gerani, M. J. Carman, and F. Crestani. Investigating learning approaches for blog post opinion retrieval. In *Proceedings of the 31th European Conference on Information Retrieval*, pages 313–324, 2009.

[45] H. Ghorbel and D. Jacot. Sentiment analysis of French movie reviews. In *Advances in Distributed Agent-Based Retrieval Tools*, volume 361 of *Studies in Computational Intelligence*, pages 97–108. Springer Berlin / Heidelberg, 2011.

[46] A.-L. Gînscă, E. Boroş, A. Iftene, D. Trandabăţ, M. Toader, M. Corîci, C.-A. Perez, and D. Cristea. Sentimatrix: multilingual sentiment analysis service. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 189–195, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[47] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 8th Conference on European chapter of the Association for Computational Linguistics*, EACL '97, pages 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[48] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur. The shark-search algorithm. An application: tailored Web site mapping. In *Proceedings of the 7th International Conference on World Wide Web*, pages 317–326, Brisbane, Australia, Apr. 1998. Elsevier Science.

[49] K. Hiroshi, N. Tetsuya, and W. Hideo. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 494–500, 2004.

[50] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International conference on World Wide Web*, WWW '07, pages 151–160, New York, NY, USA, 2007. ACM.

[51] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[52] N. Jindal and B. Liu. Review spam detection. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1189–1190, New York, NY, USA, 2007. ACM.

[53] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining*, WSDM '08, pages 219–230, New York, NY, USA, 2008. ACM.

[54] J. Johnson, K. Tsioutsiouliklis, and C. L. Giles. Evolving strategies for focused web crawling. In *Proceedings of the 20th International Conference Machine Learning*, pages 298–305, 2003.

[55] S. D. Kamvar and J. Harris. We feel fine and searching the emotional web. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 117–126, 2011.

[56] M. Kaya, G. Fidan, and I. H. Toroslu. Sentiment analysis of Turkish political news. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 174–180, Washington, DC, USA, 2012. IEEE Computer Society.

[57] A. Kennedy and D. Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

[58] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[59] M. Koster. A standard for robot exclusion, http://www.robotstxt.org/orig.html, last visit on September 1, 2013.

[60] O. Kucuktunc, B. B. Cambazoglu, I. Weber, and H. Ferhatosmanoglu. A large-scale sentiment analysis for Yahoo! answers. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 633–642, New York, NY, USA, 2012. ACM.

[61] K. Lerman, S. Blair-Goldensohn, and R. McDonald. Sentiment summarization: Evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 514–522, 2009.

[62] J. Li, K. Furuse, and K. Yamaguchi. Focused crawling by exploiting anchor text using decision tree. In *Special interest tracks and posters of the 14th International Conference on World Wide Web*, WWW '05, pages 1190–1191, New York, NY, USA, 2005. ACM.

[63] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Springer, 2011.

[64] H. Liu, E. Milios, and J. Janssen. Probabilistic models for focused web crawling. In *Proceedings of the 6th ACM International Workshop on Web Information and Data Management*, pages 16–22, 2004.

[65] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008.

[66] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Building domain-specific search engines with machine learning techniques. In *Proceedings of AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, pages 28–39, 1999.

[67] F. Menczer and R. Belew. Adaptive retrieval agents: Internalizing local context and scaling up to the web. *Machine Learning*, 39(2/3):203–242, 2000.

[68] F. Menczer, G. Pant, P. Srinivasan, and M. E. Ruiz. Evaluating topic-driven web crawlers. In *Proceedings of the 24th Annual International ACM/SIGIR Conference*, 2001.

[69] G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[70] M. Najork. Web crawler architecture. In *Encyclopedia of Database Systems*, pages 3462–3465. 2009.

[71] A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke. Predicting IMDB movie ratings using social media. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, ECIR'12, pages 503–507, Berlin, Heidelberg, 2012. Springer-Verlag.

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

[73] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of ACL-02 Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002.

[74] G. Pant and P. Srinivasan. Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems*, 23(4):430–462, Oct. 2005.

[75] B. Pinkerton. Finding what people want: Experiences with the webcrawler. In *Proceedings of the 2nd International World Wide Web Conference*, 1994.

[76] P. N. Priyatam, S. R. Vaddepally, and V. Varma. Domain specific search in Indian languages. In *Proceedings of the 1st workshop on Information and knowledge management for developing region*, IKM4DR '12, pages 23–30, New York, NY, USA, 2012. ACM.

[77] J. Rennie and A. McCallum. Using reinforcement learning to spider the web efficiently. In *Proceedings of the 16th International Conference on Machine Learning*, ICML '99, pages 335–343, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[78] G. Saffar. Crawler4j open source web crawler for Java, http://code.google.com/p/crawler4j/, last visit on September 1, 2013.

[79] N. Shuyo. Language detection library for Java, http://code.google.com/p/language-detection/, last visit on September 1, 2013.

[80] M. Sokolova and G. Lapalme. Verbs speak loud: Verb categories in learning polarity and strength of opinions. In *Proceedings of the Canadian Society for Computational Studies of Intelligence, 21st Conference on Advances in Artificial*

Intelligence, Canadian AI'08, pages 320–331, Berlin, Heidelberg, 2008. Springer-Verlag.

[81] P. Srinivasan, J. Mitchell, O. Bodenreider, G. Pant, and F. Menczer. Web crawling agents for retrieving biomedical information. In *Proceedings of Workshop on Agents in Bioinformatics*, 2002.

[82] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, June 2011.

[83] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[84] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[85] C. Toprak, N. Jakob, and I. Gurevych. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 575–584, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[86] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[87] G. Vural, B. B. Cambazoglu, and P. Senkul. Sentiment-focused web crawling. In *Proceedings of the 21th ACM International Conference Information and Knowledge Management*, pages 2020–2024, 2012.

[88] G. Vural, B. B. Cambazoglu, P. Senkul, and O. Tokgoz. A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish. In *Computer and Information Sciences III - 27th International Symposium on Computer and Information Sciences*, pages 437–445, 2013.

[89] X. Wan. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553–561, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[90] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 427–434, 2003.

[91] M. Yuvarani, N. Iyengar, and A. Kannan. LSCrawler: A framework for an enhanced focused web crawler based on link semantics. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 794–800, 2006.

[92] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, and W. Zuo. Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12):2474–2487, 2009.

[93] W. Zhang, C. Yu, and W. Meng. Opinion retrieval from blogs. In *Proceedings of the 16th ACM International Conference on Information and Knowledge Management*, pages 831–840, 2007.

APPENDIX A

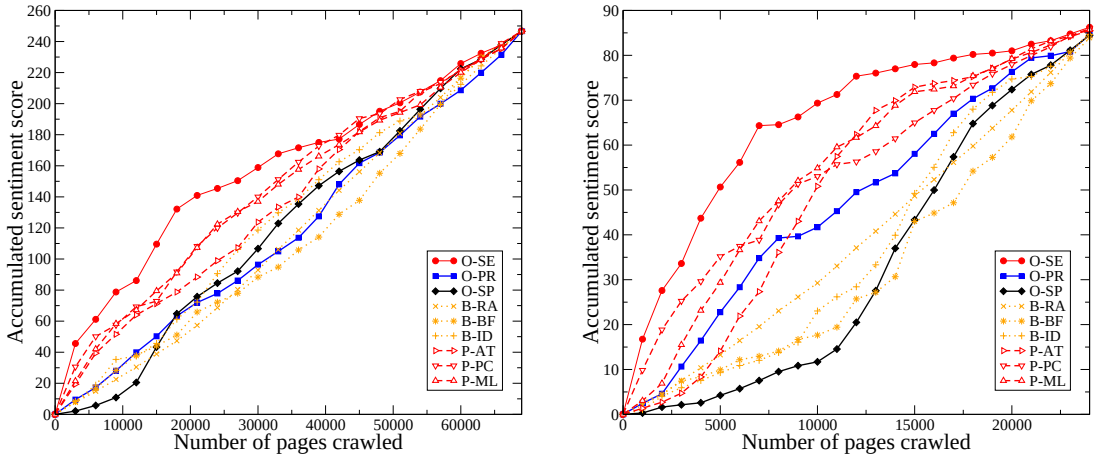
ADDITIONAL EXPERIMENTS FOR SENTIMENT-FOCUSED WEB CRAWLING

This appendix reports the experiment results given in [87]. The experiments are executed on a random and smaller subset of TREC Category B, with a size of 1,185,385 web pages. The subset has the following statistics:

- A page contains 696.0 words (247.1 unique words) and 16.0 sentences on average.
- A page has 63.1 outgoing and 14.9 incoming links on average.
- A page has a size of 30,556 and 3,228 bytes before and after parsing the HTML tags, respectively.
- About 15.8% do not link to any other page whereas about 26.1% do not receive a link from any other page.
- About 53.3% fall into the spam category.
- Majority of the pages belong to `com` domain with rate 73.5%. The other significant domains for pages are `org` (10.3%), `edu` (6.8%), `net` (6.3%), `gov` (1.5%), and `info` (1.2%).
- Similar to page domains, majority of the outgoing links belong to `com` domain with rate 79.2%. The other significant domains for links are `org` (8.7%), `net` (6.3%), `edu` (3.5%), `gov` (1.0%), and `info` (0.9%).
- Sentiment score, spam score and PageRank distributions are shown in Table A.1.

TableA.1: Some distribution statistics for sample dataset

Distribution	Min.	Avg.	Max.	σ
Sentiment score	0.000	0.004	0.276	0.009
Spam score	0.000	60.983	99.000	28.208
PageRank	0.150	0.176	90.689	0.242



(a) Sentiment accumulation without spam filtering

(b) Sentiment accumulation with spam filtering

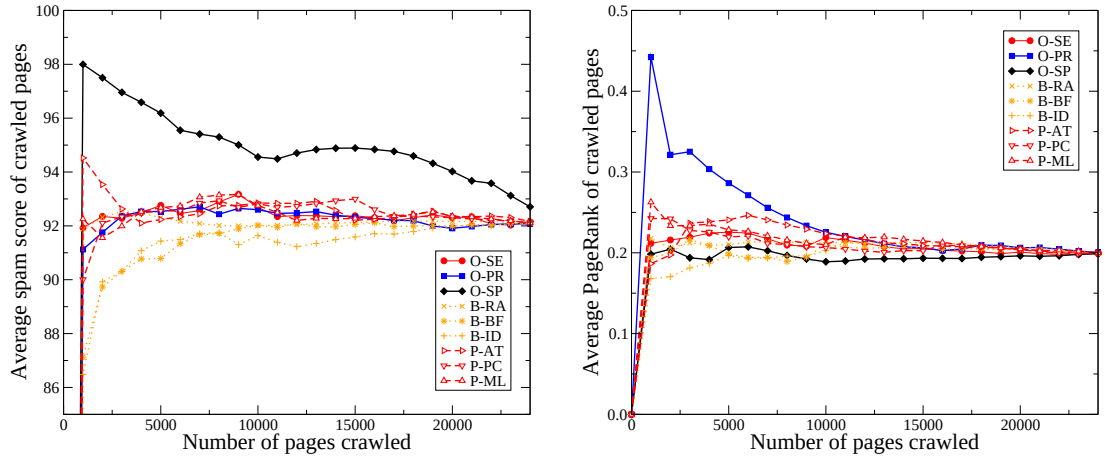
Figure A.1: Sentimentality accumulation while pages are crawled from seeds with highest outgoing links.

We simulate the crawlers with parameter combination `BP-WS-Adj`, as concluded in Chapter 3.3 to the best one for ground-truth. In all simulations, the seed pages are set as the first 1,000 pages with the highest number outgoing links.

The observations of the total amount of sentimentality accumulated during the crawling process are shown in Figure A.1(a) and Figure A.1(b). The main difference between these figures is the spam filtering strategy applied, where spam filtering is off and on respectively. When spam filtering is on, any fetched page assumed as spam is ignored by the crawler. For our sample set, starting from the selected seeds the crawler can access about 24K and 69K pages when spam filter is on and off respectively.

According to Figure A.1(a), as expected, the oracle crawler (`O-SE`) achieves the best performance. It is important to note that oracle crawler accumulates almost half of the sentimentality available in the web page sample after crawling only quarter of the accessible pages. All the proposed sentiment-focused crawlers (`P-PC, P-ML, P-AT`) perform well approaching to performance of the oracle crawler (`O-SE`). The results of this simulation show that *i)* the links in sentimental pages are likely to lead to other sentimental pages, *ii)* the sentimentality of a page can be accurately estimated to a certain degree without having its content, and *iii)* there is no significant correlation between the PageRank and the sentimentality of a web page.

As stated above, Figure A.1(b) shows the total amount of sentimentality accumulated during the crawling process when the spam pages are ignored. Unsurprisingly, `O-SE` is the best performing strategy. `O-SE` accumulates almost 65% of the sentimentality available in the web page sample again after crawling only quarter of the accessible pages. All the proposed sentiment-focused crawlers (`P-PC, P-ML, P-AT`) perform again well with variable performances at different stages. This variance could be a re-



(a) Average spam score accumulation with spam filtering

(b) Average PageRank accumulation with spam filtering

Figure A.2: Results while pages are crawled from seeds with highest outgoing links.

sult of disconnectedness of links due to spam page ignorance. An interesting finding of this experiment is that oracle crawler with PageRank (`O-PR`) performs well, especially at early stages. Similar to Figure A.1(a), lowest depth baseline crawler (`B-BF`) has the poorest performance.

Fig. A.2(a) displays the variation of the average spam score of the downloaded pages when the spam filter is present. As expected, the spam oracle `O-SP` quickly forms a collection with high spam scores. As seen, the sentiment-focused web crawling techniques do not create a negative bias since they do not lead to a significant increase in the spam rate of the crawled pages.

Fig. A.2(b) shows the average PageRank values of the downloaded pages when the spam filter is present. As expected, the PageRank oracle `O-PR` quickly forms a collection with high PageRank scores. The sentiment-focused crawling techniques also perform relatively good, which implies that the sentiment-focused web crawling techniques do not create a negative bias in terms of PageRank.

It can be concluded that all these results are consistent with the results reported in Section 3.6, and with the outstanding performances of proposed strategies, sentiment-focused web crawling is feasible.

APPENDIX B

EXAMPLE TURKISH DATA

B.1 Movie Review Data

The following subsections present one positive and one negative movie review respectively.

B.1.1 Positive Movie Review

"harika tek kelimeyle harika bir film oyuncular filmi izlettirmekle kalmıyor adeta filmi yaşattırıyorlar"

B.1.2 Negative Movie Review

"Malesef buyuk hayal kirikligi.. shyamalan'in en kötü filmi bence... biraz ask biraz dram filmde baska bir sey yok.. zaman ve pazar kaybi bence..... fragmana aldanmamak lazimmis....."

B.2 Hotel Review Data

The following subsections present one positive and one negative hotel review respectively.

B.2.1 Positive Hotel Review

"Herşey harikaydı. Geçen yıl nisan ayında çok uzak olmaması için Kuşadasında tatil yapalım dedik.Ve Ephesia yı bulduk resimlerinden çok güzel bir otel olduğu belliydi ama resimlerden çok daha iyi çıktı ki eminim yaz sezonunda çok daha iyidir."

B.2.2 Negative Hotel Review

“Eğer gidip yatıcam diyorsanız tavsiye ederim yalnız bu paraya eğlencede olsun diyorsanız tavsiye etmem. Ücret kalite aşırı dengesiz. Organızasyonda sıkıntı var. Nerde ne var belli değil. Hizmet kalitesi yemekler içecekler insanların davranışları iyi.”

B.3 Political News Data

The following subsections present one positive and one negative political news respectively.

B.3.1 Positive Political News

“kendine haber kanalı diyen onca televizyonun ; tabiri caizse ”namusu kurtardığı” bir iki programı var. bunlardan biri, hiç şüphe yok ki ahmet hakan’ın ”tarafsız bölge”si. geçen akşam türkiye’nin kısır siyaseti tartışılırken, bunca zamandır ilk kez ilgimi çeken bir ses, nihayet yükseldi. ateş ilyas başsoy. genellikle böyle övgüler pek bir manasızdır. ve ben böylesi bir girizgahla arkadaşı övüyorsam inanın bir şey vardır. bunun onun tanımadığım şahsı, ekrandaki pozitif diliyle ilgisi yok. iletişim arenası o denli çölleşmiş vaziyetteki birikiminin haklı kibrini yüzümüze bağırarak kusmayacak nazik genç bilgelere susadık. o sebeple, hiç tanışmamakla birlikte samimi tespitleri, bugüne kadar okuduğum, dinlediğim birçok siyasetçi, akademisyen, gazeteci veya aydından öte. kitabını okumadım. ”akp neden kazanır? chp neden kaybeder?” ötesi var mı? derhal okuyacağım. ama programda dinlerken genel çerçevesine dair az çok fikir sahibi oldum. şimdi gelelim selim türkhan’a. kaybedenler kulübü’nde çok sevdiğim bir replik vardı. ”kim ulan bu erol egemen?” öyle biri yoktu ama bir semboldü. selim böyle bir karakter. selim türkhan’ı ateş’ten kısaca dinleyelim; selim türkhan, bir esnaf (veya memur veya işçi veya patron veya işsiz). karısı öğretmen emeklisi. büyük kızı üniversitede okuyor, oğlu da üniversite sınavlarına hazırlanıyor. (her biri için ayrı ayrı ’veya’lar üretilebilir.) selim türkhan’ın her cümlesi son derece emin tonlarla çıkıyorsa da, yine her cümle siyaseten akıl almaz çelişkiler içeriyor. bunların selim türkhan için hiç önemi yok. o aynı anda hem milliyetçi, hem dindar, hem modern, hem laik ve en önemlisi hem de bunların hiçbiri olabilir. selim türkhan türkiye’nin orta sınıfı, zengini veya yoksulu olabilir. selim türkhan türk, çerkez, gürcü, kürt, pomak olabilir. selim türkhan eşimiz, dostumuz, kardeşimiz, ebeveynimiz olabilir. selim türkhan, akp’nin büyük oranda ikna ettiği ve diğer partilerin hemen hiç ikna edemediği, siyasetten ayrışmış ’kendi halinde’ erkek ve kadınların sembolü. selim türkhanlara ’kararsız’ diyorlar. ben bunu yanlış buluyor ve onları ’siyasetsiz’ diye tanımlıyorum. siyasi söylemler, siyasi kavgalar, siyasi doğrular veya yanlışlar selim türkhan için belirleyici değil. o bu konuları fazlaca dinlemiyor ve taraf olmuyor. türkiye seçmeninin yaklaşık %30’u selim türkhan’dan oluşuyor. ateş, selim’i uzun uzun anlatıyor. ama siz onu tanıdınız. ben de tanıdım.

chp, neden kaybediyor sorunsalının dna'sını selim'deki türkiye'de arayın. ben buldum. ve daha çok konuşacağız " (Serdar Akinan, Akşam, 01.02.2012)

B.3.2 Negative Political News

"Şurası çok açık görülüyor... Mustafa Balbay'ın içeri tıkılması, gazetecilerin çoğunu mutlu etti. Çünkü... Mustafa Balbay gibi bir gazeteci, mesela restorana girdiği zaman, yan masada oturan insanlar ona selam verir, fikirlerine katılmasa bile saygı gösterir. Ama, Balbay'ın içeri tıkılmasından mutlu olan gazeteci kılıklı tipleri gören insanlar, "Bak bu şerefsiz de buraya gelmiş" derler. Çünkü... İnsanlar biliyor, Soros'tan para alanları, AB'cileri, Washingtoncuları, Ali Kemalleri, besleme yalakaları, imam kökenli Reina fırıldaklarını, bir yandan generallerin kıçını, bir yandan tarikatçıların eteğini öpenleri, liboşları... Biliyor insanlar. Tanıyor. O nedenle... Mustafa Balbay'ı yolda yürürken görürlerse, elini sıkıp teşekkür ederler, "İyi ki varsın" falan derler, fotoğraf çektirirler. Öbürlerini gördüklerinde, tükürür gibi bakarlar. Tükürürler de bazen. Anlatıyorlar. O nedenle... Mustafa Balbay gibilerini davet eder üniversite öğrencileri, gelsin konuşsun diye... Öbürlerinin ise, bakmayın "Türbanlılar niye giremiyor" diye yazıp çizdiklerine, aslında kendileri giremiyor o üniversitelere! İyi oldu, iyi. Hepsi birarada olmuyor. Mustafa Balbay gibiler içeri girecek ki, öbürleri dışarda gezebilsin, kıyaslanmadan." (Yılmaz Özdil, Hürriyet, 04.07.2008)

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Vural, Avni Güral
Nationality: Turkish (TC)
Date and Place of Birth: 4 August 1976, Ankara
Marital Status: Married
Phone: +90 532 384 30 74
e-Mail: guralvural@yahoo.com

EDUCATION

Degree	Institution	Year of Graduation
MS	University of Manchester – Computer Science	2000
BS	Bilkent University – Computer Engineering	1999
High School	Ankara Atatürk Anatolian High School	1994

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2012 – present	MilSOFT	Senior Expert Software Engineer
2008 – 2012	MilSOFT ICT	Senior Software Engineer
2000 – 2008	MilSOFT	Software Engineer
1998	Procter&Gamble	Summer Intern
1997	IBM	Summer Intern

PUBLICATIONS

International Conference Publications

1. Gural Vural, B. Barla Cambazoglu, and Pinar Senkul. Sentiment-focused web crawling. In *Proceedings of the 21th ACM Internation Conference Information and Knowledge Management*, pages 2020–2024, 2012.

2. Gural Vural, B. Barla Cambazoglu, Pinar Senkul, and Ozge Tokgoz. A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish. In *Computer and Information Sciences III - 27th International Symposium on Computer and Information Sciences*, pages 437–445, 2012.

National Conference Publications

1. Turgay Celik, Gural Vural, and Serdar Baklan. Canlı, sanal, yapısal simülasyon sistemlerinin entegrasyonuna yönelik örnek bir çalışma. *5. Ulusal Savunma Uygulamaları Modelleme ve Simülasyon Konferansı*, pages 301–310, 2013.

2. Turgay Celik and Gural Vural. Savunma sanayii projelerinde Java dili ve teknolojilerinin kullanımının değerlendirilmesi. *6. Savunma Teknolojileri Kongresi*, 2012.