ISSN 1870-4069

A Document Profile for Improving Information Retrieval Systems

Antonio Guillén, Yoan Gutiérrez, Rafael Muñoz

University of Alicante, San Vicente del Raspeig,
Department of Software and Computing Systems, Alicante, Spain

{aguillen,ygutierrez,rafael}@dlsi.ua.es

https://www.dlsi.ua.es

Abstract. In recent year the great popularity that enjoys mobile technologies has led most users to become consumers and producers of information on the network. Many studies speak about this phenomenon as an activity that is capable of doubling or tripling existing content on an annual basis. The huge amount of information makes the current user oriented systems, like retrieval systems, recommender systems and others, become less efficient, especially when users require specific information and answers according to their needs and preferences. These facts make necessary to equip these systems with proper Natural Language technologies able to provide the information that users demand adapted to each context and content type. In this article, it is presented a study of some Natural Language Processing technologies that can be useful to facilitate the proper identification of documents according to the user needs. For this purpose, it is designed a document profile that will be able to represent semantic meta-data extracted from documents. The research is basically focused on the study of different language technologies in order to support the creation this novel document profile proposal from semantic perspectives.

Keywords. Retrieval system, NLP technologies, document profile, meta-data, web searching.

1 Introduction

The Internet provides large amounts of information through many types of documents. Users require an easy way to filter these documents to find out the most appropriate documents for their interests, capabilities and needs. These documents also can be retrieved by other entities for market studies, classify and index documents, detect fake information or illegal activities on the Web. These aspects can be treated by the study of NLP areas, tasks, methods and tools.

In this paper, is presented a novel document profile proposal for improving information retrieval systems. Also, is presented a study to determine which NLP

technologies are the most suitable for this purpose. In this study, is addressed which technologies are available currently, estimate its automation and reliability degree, the problems that can be found on applying them and the most appropriate document's type. Also, is described this novel document profile proposal and is presented a selection of documents to be treated in the proposal. A document can be defined as a short or long unity of information, principally obtained from websites (user posts, press articles, comments, reviews, etc.). The aim of this study is to investigate the different features that can be extracted by means NLP technologies able to provide enough information for setting up useful meta-data. The main purpose of this meta-data is improving retrieval systems and search engines to get better search results. Also, another important purpose is to improve recommender systems through common meta-data between documents.

The paper is organised as follows: Section 2 shows the related work of this research. Section 3 describes the novel document profile proposal. Section 4 presents an example of document profile with a real document and a possible usage. Section 5 addresses a study of selected NLP tasks and technologies which serve for supporting the proposal. Section 6 exposes the conclusions and future works.

2 Related Work

Two areas closely related to Document Profiling are User Profiling [1] and Recommender Systems [2]. Although those research areas differ from Document Profiling we can find some features useful to be reused and incorporated into this work. For instance, author identification methods help to reveal personal information (i.e. age, gender) about document's author. In the case of reusing Recommender Systems research technologies, there are interesting researches for our work. Paper [3] propose an approach for constructing user profiles based on the interaction and user behaviour on the Internet for news recommendation, this profile is relevant to our work because is focused on news documents and includes information about time expressions, location, and topic tags.

There are other works related to the document profile proposal. Paper [4] justifies the importance of this proposal for Information Retrieval. This work emphasises the search quality by using meta-data information. Paper [5] uses the term *document profile* to refer to N-Gram frequencies for text categorisation. Paper [6] explains a review analyser that extracts product reviews and analyse its information. As result sentiment polarity classifications are extracted. Other related works are focused on improving retrieval systems [7] tough efficient Information Retrieval techniques and algorithms. It is interesting because it reveals some useful web searching techniques and algorithms.

3 Document Profile Proposal

The proposal mainly consists of designing a document profile able to represent different features extracted once NLP technologies are applied on documents,

assisted or not by humans. As can be seen, there are many features that can only be extracted automatically from documents by using NLP technologies. This work pretends to unify most of these NLP technologies as a whole able to characterise documents from different points of view.

3.1 Document Selection

As a first approach, this work will be focused on English documents from the Internet. This is due to NLP technologies have been mostly developed to cover this language and which makes easier to find out NLP tools. Nevertheless, the overall strategy is still valid for other languages. The principal documents addressed are those mostly available on the Internet[1]. Those are press media, social posts, product/service reviews, blogs or personal websites, academic/science papers, tutorials or instruction manuals, etc. The complete documents' schema and meta-data properties are shown in this source[2]. Is defined two classes of documents (leaf and conceptual), leaf document represents concrete documents that can be obtained the profile, and the conceptual document is defined as group documents that share common features.

Leaf documents represent Internet content such as blogs, tweets, news, etc. This work is focused on these documents because is daily used by users on the Internet, and considered in many NLP corpora.

Next, is described some of these documents, which can be represented by the leafs:

Press Document: A press document consists of a digital press document provided by any press media (TV, written, etc.). Usually, this document is a large text which use to include information about entities like dates, persons, places, institutions, etc. Some examples of press documents are on-line versions of newspapers such as The New York Times; The Times; The Washington Post; The Economist; or news channels like CNN, BBC World News etc.

Social Post: A social post is a content coming from any social network or forum. Usually, it is a short informal text written by an Internet user. Some examples of social post documents can be found in popular social networks[1] like Twitter, Facebook, Instagram, etc.

Product/Service Review: A product review is an opinion expressed by a consumer about a product, service, video, etc. Usually, it is a short informal text. Some examples of product reviews can be found on e-Commerce websites like Amazon[3], video-sharing websites like Youtube, etc.

Blog: A blog document is an informal diary-style text entry about an author ideas. Can be a large or short informal text written by a user. Many examples of blogs are found on the Internet about many subjects or topics. Some blogs[4] examples are The Huffington Post, TMZ, Business Insider, etc.

[1] http://www.quantcast.com/top-sites
[2] http://ow.ly/pTaR30dWTj7
[3] https://www.amazon.com/review/hall-of-fame
[4] http://www.ebizmba.com/articles/blogs

Personal Web: A personal web document is a personal website or biography in which a person describes synopses of her/his life. Usually, it is a large formal text about a person in which it can be identified many clues about her/his life like dates, places, companies, other persons, etc. Also, it can be considered a personal web a general purpose webs[1] like commercial websites, corporations, brands, products or services.

Academic Document: An academic document is any digital content used to the learning process, i.e. digital books, slides, etc. Usually, it is a large formal text used for whatever level teaching.

Instruction Manual: An instruction manual is a large formal text for describing technical procedures or how instructions for using products or services.

Tutorial: A tutorial document is a set of instructions to learn some task. Usually, it is a large informal text written by users.

Scientific Document: A scientific document is whatever scientific environment material like articles, slide-shows, etc. Usually, it is a large formal text used for whatever level teaching.

Literary Document: A literary document is a text content usually represented by E-Books. In this work, it will be considered only a portion of a literary work because a complete work maybe is very large to analyse. Therefore, a literary document represents a short formal text written by an author.

Technical Document: A technical document represents textual contents that describe professional or commercial procedures (corporate websites, descriptions of products or services), medicine and health (medical prescriptions, pharmaceutical leaflets), public administration (public calls, court judgements, tenders), among others.

3.2 NLP Tasks Selection

In this research, a selection of NLP tasks[5] for designing the profiling meta-data properties is carried out. This list has been taken into account to be included in this novel document profile proposal. This list has been taken into consideration to be included in this novel document profile proposal. At this way, it can be obtained a long quantity of meta-data contributing to filling document profiles. As can be seen, is chosen some of the most active NLP tasks nowadays. So, the main goal at this stage is to find out friendly access NLP technologies (i.e. tools, APIs, demos, etc) to test and simulate this proposal in a real scenario. At this way, it would be demonstrated the viability of this proposal, not just providing a document profiling scheme.

3.3 Meta-data Properties

The meta-data properties defined for our proposal can be found in the following source[6]. Notice, that each of these properties indicates the NLP technology acronym from which it is obtained. Is included a brief description of each meta-data property.

[5] http://ow.ly/MPO130e24Jr
[6] http://ow.ly/tQos30dNVcS

3.4 Document Profiling Algorithm

Is defined an algorithm to generate document profiles using the available NLP technologies. Algorithm 1 shows the steps to generate a profile from a Web document.

Algorithm 1: Document Profiling

```
program GenerateProfile
   {Require: url};
   begin

     cont := getContent(url);
     sum := getSummary(cont);
     type := getType(cont);
     profile := newProfile(cont,sum,type);
     listProperties := getProperties(type);
     foreach (prop in listProperties) do
       nlpTech := prop.getNlpTech();
       value := nlpTech(cont,summ);
       profile.add(prop,value);
     endfor

   end.
```

The steps are specified as follows: Web document content is extracted from url. A short version of the document using the summarising technology is generated. Detection of the document type. In the first approach, this detection will be manual. It is intended to do a classification system for auto-detection using the document's schema proposed. Generate an initial profile from scratch, only including the meta-data obtained currently (i.e. content, summary and type). Get the specific meta-data properties for this document type. Iterate on meta-data property list. Get related NLP technology of meta-data property. Invoke NLP technology with the content or summarised document (depends if require a large or short version document). Add result value obtained from NLP technology to the profile.

4 Document Profile Example and Usage

In order to figure out how would be a document profile, we have prepared an example considering a document taken from the CNN website. The document is a real news article about Everest's climber George Mallory. He's the first person who tried climb to summit the Everest. He was disappeared and his body was found 75 years later. Expert people tried to discover if he reached the top of Mount Everest or not. The complete news article it's available in CNN

website[7]. The procedure follows as to describe the algorithm 1, obtaining the next document profile[8].

A brief description of the features set is commented following. (i) The identification *001* is auto-generated. (ii) The document type detected is *Press Document*. (iii) *Content*, *Title* and *Date* meta-data properties can be extracted directly from the article, and the source corresponds to the web URL of article. (iv) The content summary is generated. (v) The topics list represents the most frequently used terms in the text. (vi) The region is obtained from the text, in this case, it talks about the north ridge of Mount Everest located in Tibet (China). (vii) About the subject areas detected can see *History* because it is a historical news article and *Sports* it talks about climbing. (viii) The language detected is English. (ix) The article does not present any rating. (x) Keywords are the more representative words and expressions of the whole text. (xi) In this article, it is not detected any ideological orientation because it only talks about a historical fact related to sports. (xii) Sentiment polarity determines that it is a positive news for people interested in history, climbing and the Mount Everest. (xiii) It is detected the emotion category *Surprise* because the man's body has been found 75 years later, something unexpected. (xiv) Some time expressions are detected in the text, these expressions have been converted into date format. (xv) The name entities detected refer persons and locations retrieved in the text. (xvi) In this article is not detected irony. (xvii) Reading complexity is *Easy*, with *Normal* textual formality and *Neutral* writing formality because the news article is from a serious press media. (xviii) The article is feasible to be read by 16 years old people over, since it is a neutral and simplified information. (xix) The age of author is predicted in the ranges of 13-17 years old or 65-100 years old. (xx) The gender is predicted as *Female*. (xxi) The press type is classified as *News Article* and its veracity is *Truthful*, taking into account the formality of press media.

One interesting use of our proposal is shown in this source[9]. Through document profile, it is possible to get related documents with common meta-data. For instance, with some meta-data could navigate among different documents that can be interesting to the user. This usage provides a great experience of knowing automatically which are the documents more appropriate to each user.

5 NLP Technologies Study

Each NLP task considered is described in this work and exposes some available technologies: tools, APIs or demos related. In some cases, only methods and algorithms had been found. Mainly works with evaluations and results are presented. In this study, the interests are mainly focused on available technology (Web service/API, programming library) with certain automation and reliability degree to be able to be incorporated in future frameworks or prototypes. The results of this study are exposed in a comparison table.

[7] http://ow.ly/7JtB305V8z6
[8] http://ow.ly/Pnmj30dNVtQ
[9] http://ow.ly/Etos30gfoH9

Text Classification: Text classification task (TC) consists of identifying the type of document by content analysis. The relevance of this task in document profiling is to determinate which type of document is analysed, depending its type different features would be extracted. For example, Dandelion API[10] categorises plain text on eight categories: business, economy, sports, etc. In this approach, the type is considered a precondition data to generate a profile, because it doesn't exist a TC technology that uses the concrete leaf document types proposed documents' schema.

Information Extraction: Information Extraction task (IE) has been addressed by the paper [8] as a way to search and obtain text on large volumes of unstructured information to filter relevant information, using regular expressions, rules and patterns. This is useful for document profiling because many complex meta-data can be obtained directly from the document. Since this work is mostly focused on Web documents there are too many tools like DEiXTo[11] that can extract information from the W3C DOM documents. The problem with these tools is the low-automation degree due to they should be re-configured.

Topic Recognition: Topic Recognition task (TR) consists of identifying topics in the text. This task is interesting to classify a document in multiple categories or topics and know the different aspects that dealt the document. TextRazor[12] is a Web tool that lists topics from a text. Also, another Web demo is Meaning Cloud[13] that offers many NLP services among which topic recognition is included.

Keyword Extraction: Keyword Extraction task (KE) is the automatic extraction of relevant terms from a document. Unlike TR, KE doesn't intend to know the different aspects that dealt the document, KE extracts terms that best describe the subject of the document. Statistical Keyword Extraction Tool (SKET) [9] is a programming library for extracting keywords from the text.

Named Entity Recognition: Named Entity Recognition task (NER) tries to locate and classify named entities according to different categories like names of persons, organizations, etc. Stanford NER [10] is a NLP technology available as a programming library and evaluated in some scenarios, domains and corpora, that offers useful NER services.

Time Expression Recognition: Time Expression Recognition task (TER) consists of obtaining temporal expressions from texts. This information is useful to extract historical facts in texts or documents. Stanford SUTime [11] is a tested technology and available to be used as a programming library. An alternative is TIPSem [12] which has been tested and available as API.

Automatic Summarization: Automatic Summarization task (AS) obtains a reduced text from a larger text content. It's interesting to obtain a short

[10] http://dandelion.eu
[11] http://dexi.io
[12] http://www.textrazor.com
[13] http://www.meaningcloud.com

version of the same document. Paper [13] presents a summarization system for various purposes and domains. This system has been evaluated and is available as API.

Domain Detection: Domain Detection task (DD) is part of Semantic Parsing. This task detects the meaning of sentence using probabilistic semantic models. Paper [14] is focused on ISR-WN which is able to detect domains or categories from different resources obtaining and using relevant semantic trees from a text.

Language Identification: Language Identification task (LI) consists of identifying the language. AlchemyLanguage[14] is a Web demo that offers many NLP services for that purpose.

Polarity Classification: Polarity Classification task (PC) is part of Sentiment Analysis. According to paper [15], this consists of determining whether a text in positive, negative or neutral. Many PC approaches are using machine and deep learning which obtains good results. Among others, are mentioned two relevant works: [16] which presents an approach using supervised statistical machine learning, and Stanford Sentiment [17] that is available as a programming library.

Emotion Detection: Emotion Detection task (ED) is part of Sentiment Analysis. This consists of identifying some emotions expressed in texts. The following set of the basic emotional categories proposed by Ekman [18]: Anger, Disgust, Fear, Joy, Sadness and Surprise. One issue of this task is the lack of annotated corpus for evaluation. ToneAnalyzer[15] is a web demo of multiple NLP APIs, including the emotion detection task with 5 Ekman categories: Anger, Disgust, Fear, Joy and Sadness.

Readability Analysis: Readability Analysis task (RA) according to paper [19], the detection of RC consists of determining documents suitable for being read by specific people age ranges. This includes determining reading difficulties or text comprehension. For addressing this task it is necessary to classify documents on the basis of different levels of reading comprehension. Different measures of readability must be selected, the work is based on the study Flesh-Kincaid Grade Level. This study tries to predict the recommended age to understand the text. The tool Readable.io[16] is a web demo for evaluating the readability level of a text using various grade levels, including Flesh-Kincaid Grade Level.

Informality Analysis: Informality Analysis task (IA) tries to detect the degree informality in texts. This task arises due to the necessity of processing in a personalised way non-traditional textual sources existing on the Internet (i.e. blogs, forums, etc.). TENOR [20] provides functionalities aligned to this task.

Age Estimation: Age Estimation task (AE) is part of the Author Profiling area. According to paper [21], AE tries to predict some aspects of authors

[14] http://alchemy-language-demo.mybluemix.net
[15] http://tone-analyzer-demo.mybluemix.net
[16] http://readable.io

like age or gender. This task will contribute to this research the discovery of various authors types depending on age ranges. Age Analyzer[17] is a web API that provides functionalities aligned to this task.

Gender Detection: Gender Detection task (GD) is strongly related to AE task, but in this case, it tries to detect the gender of text's author. Gender Analyzer[18] is appropriated to this task as has been before mentioned.

Irony Detection: Irony Detection task (ID) tries to detect if a literal message has an opposite meaning, without a negation marker. The difficulty resides the absence of face-to-face contact and vocal intonation. In the automatic detection of irony is used sentiment analysis, information extraction or decision making to obtain textual features for recognising irony. Paper [22] presents a research for irony detection in Twitter short documents, using the tasks mentioned above.

Ideology Detection: Ideology Detection task (IDD) tries to detect the ideology orientation expressed in a text content based on a set of opinions or beliefs. Usually, this refers to a set of political beliefs or a set of ideas that characterise a particular culture. Paper [23] presents a research work where political ideology orientation is detected using neural network technologies.

5.1 Automation and Reliability Study

In the present study, the high-reliability degree is defined as technologies that have an evaluation with high scores of performance. Similarly, the high-automation degree is defined as the type of technology easier to implement or use in each case. For example, Web Services, Java or Python libraries, Algorithms, Web application/demo and Desktop tools. Web Services or online APIs present a very-high-automation degree because it is easy to use them in frameworks or meta-tools developments. Java or Python libraries have a high-automation degree, because it is easy to include them in frameworks or meta-tools develops, but sometimes it is needed to proceed with an adaptation stage of the target programming framework. Algorithms present a medium-automation degree because it should be considered the efforts of reproducing them. Web application/demo tools have a low-automation degree because initially, it is difficult to automatically include them, however, an alternative is using web crawling.

Most experimental approaches that make use of Web application/demo tools apply semi-automatic procedures. Desktop tools have a very-low-automation degree because it is very difficult to include them in frameworks or meta-tools develops, most of the time the procedures are performed by hand. Table 1 shows a detailed comparison at respect. The automation of document profiling procedures, such as this study reveals, is supported by the reliability degree presented in Table 1. The performing scores have been taken from the bibliography before cited, being them the best technologies found in the state of the art. The technologies where is set "Not found", probably it is because they represent

[17] http://ageanalyzer.com
[18] http://www.genderanalyzer.com

Table 1. NLP technologies comparisons.

NLP Technology	NLP task	Measure	Score	Type
Dandelion	TC	Not found	Not found	Web demo
DEiXTo	IE	Not found	Not found	Desktop
TextRazor	TR	Not found	Not found	Web demo
MeaningCloud	TR	Not found	Not found	Web demo
SKET	KE	F1	0.7	Java library
Stanford SUTime	TER	F1	0.92	Java library
TIPSem	TER	F1	0.85	Web service
Stanford NER	NER	F1	0.8876	Java library
Summarise	AS	F1	0.57797	Web service
ISR-WN	DD	F1	0.52	Web service
Alchemy Language	LI	Not found	Not found	Web demo
Kiritchenko et al. (2014)	PC	F1	0.855	Algorithm
Stanford Sentiment	PC	Accuracy	0.854	Java library
Zhang et al. (2017)	ED	F1	0.56	Algorithm
Tone Analyzer	ED	Not found	Not found	Web demo
readable.io	RA	Not found	Not found	Web demo
TENOR	IA	Not found	Not found	Algorithm
Age Analyzer	AE	Not found	Not found	Web service
Gender Analyzer	GD	Accuracy	0.95	Web service
Reyes et al. (2013)	ID	F1	0.76	Algorithm
Iyyer et al. (2014)	IDD	Accuracy	0.702	Algorithm

developments related to companies or private research. However, it is interesting to study them for future evaluations. At respect to Web demos, these offer limited NLP services which could be possible resolve through business registration or payment of services. Regarding the type "Algorithm", in some cases, this one is not available in the reference papers. Nevertheless, it is considered because the authors could be contacted in some way.

The study revealed that many NLP technologies are interesting to this aim, however, many of them are difficult to be reused. This depends on the licenses of use, visibility, replicability of the algorithms, etc.

6 Conclusion and Future Work

In this paper is presented the study of useful NLP technologies to automate the process of building document's profiles. Clearly, knowing the difficulties found to reuse the NLP technologies will help us to be more focused on considering those technologies with a high-automation degree instead of a high-reliability degree. As result, in this paper is demonstrated many different NLP technologies can converge all in a unique ecosystem. It can improve retrieval systems and even be able to recommend documents to users. As future work is planned to create a dataset for further supporting the creation and evaluation of document profile. In addition, the study of the results of evaluating different types of document

profiles (i.e. social, press, book, etc.) will be included in our research agenda. This study could include the adaptation of some NLP technologies at the proposed work, for instance, a new classification system using the documents' schema.

Acknowledgments. This work is funded by a FPU predoctoral fellowship granted by University of Alicante (UAFPU2015-5999), and has also been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government, Ministerio de Educación, Cultura y Deporte and Ayudas Fundación BBVA a equipos de investigación científica 2016 through the projects TIN2015-65100-R, TIN2015-65136-C2-2-R, PROMETEOII/2014/001, GRE16-01: "Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet" and ASAP.

References

1. Sapkota, U., Bethard, S., y Gómez, M.M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. – Hum. Lang. Technol. (NAACL HLT 2015). 93–102 (2015). doi:10.3115/v1/N15–1010
2. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowledge-Based Syst. 46, 109–132 (2013). doi:10.1016/j.knosys.2013.03.012
3. Gulla, J.A., Fidjestøl, A.D., Su, X., Castejon, H.: Implicit User Profiling in News Recommender Systems. Int. Conf. Web Inf. Syst. Technol. 185–192 (2014). doi:10.5220/0004860801850192
4. Usbeck, R.: Combining Linked Data and Statistical Information Retrieval (2014)
5. Cavnar, W.B., Trenkle, J.M., Mi, A.A.: N-Gram–Based Text Categorization. Proc. SDAIR–94, 3rd Annu. Symp. Doc. Anal. Inf. Retr. 161–175 (1994). doi:10.1.1.53.9367
6. Kshirsagar, A.A., Deshkar, P.A.: Review analyzer analysis of product reviews on WEKA classifiers. In: ICIIECS 2015 – 2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems (2015)
7. Hajeer, S.I., Ismail, R.M., Badr, N.L., Tolba, M.F.: An adaptive information retrieval system for efficient web searching. Commun. Comput. Inf. Sci. 488, 472–482 (2014)
8. Vila, K., Fernández, A., Gómez, J.M., Ferrández, A., Díaz, J.: Noise-tolerance feasibility for restricted-domain Information Retrieval systems. Data Knowl. Eng. 86, 276–294 (2013). doi:10.1016/j.datak.2013.02.002
9. Rossi, R.G., Marcacini, R.M., Oliveira Rezende, S.: Analysis of domain independent statistical keyword extraction methods for incremental clustering. J. Brazilian Soc. Comput. Intell. 12, 17–37 (2014)
10. Finkel, J., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. 43rd Annu. Meet. 363–370 (2005). doi:10.3115/1219840.1219885
11. Chang, A.X., Manning, C.D.: SUTime: A library for recognizing and normalizing time expressions. LREC. 3735–3740 (2012). doi:10.1017/CBO9781107415324.004
12. Llorens, H., Saquete, E., Navarro, B.: Temporal expression identification based on semantic roles (2009)

13. Alcón, O., Lloret, E.: Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de análisis de componentes principales para la generación de resúmenes multilingües (2015)
14. Gutiérrez, Y., Vázquez, S., Montoyo, A.: A semantic framework for textual data enrichment. Expert Syst. Appl. 57, 248–269 (2016). doi:10.1016/j.eswa.2016.03.048
15. Mohammad, S.M.: Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In: Emotion Measurement. pp. 201–237 (2016)
16. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. J. Artif. Intell. Res. 50, 723–762 (2014). doi:10.1613/jair.4272
17. Socher, R., Perelygin, A., Wu, J.: Recursive deep models for semantic compositionality over a sentiment treebank. Proc. 1631–1642 (2013). doi:10.1371/journal.pone.0073791
18. Ekman, P.: Basic Emotions. In: Handbook of Cognition and Emotion. pp. 45–60 (2005)
19. Valdivia, M.T.M., Cámara, E.M., Barbu, E., López, L.A.U., Moreda, P., Lloret, E.: Proyecto FIRST (flexible interactive reading support tool): Desarrollo de una herramienta para ayudar a personas con autismo mediante la simplificación de textos. Proces. Leng. Nat. 53, 143–146 (2014)
20. Mosquera, A., Moreda, P.: TENOR: A lexical normalisation tool for spanish web 2.0 texts. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 535–542 (2012)
21. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Inf. Process. Manag. 52, 73–92 (2016). doi:10.1016/j.ipm.2015.06.003
22. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in Twitter. Lang. Resour. Eval. 47, 239–268 (2013). doi:10.1007/s10579–012–9196-x
23. Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P.: Political Ideology Detection Using Recursive Neural Networks. Acl-2014. 1113–1122 (2014). doi:10.1017/CBO9781107415324.004