Peer-reviewed version available at *Data* **2019**, *4*, 18; doi:10.3390/data4010018

Article

Towards identifying author confidence in biomedical articles

Mihaela Onofrei [2], Diana Trandabăț [1], Daniela Gîfu[1,2,3]*

[1] Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania; {dtrandabat; daniela.gifu}@info.uaic.ro;
[2] Institute of Computer Science, Romanian Academy – Iasi branch, Romania, {mihaela.onofrei; daniela.gifu}@iit.academiaromana-is.ro;
[2] Cognos Business Consulting S.R.L., 7, Iuliu Maniu Blvd, 061072, Bucharest, Romania;
* Correspondence: daniela.gifu73@gmail.com; Tel.: +40-742-050-673

Abstract: In an era when medical literature is increasing daily, researchers in biomedical and clinical areas have joined efforts with language engineers to analyze large amount of biomedical and molecular biology literature (such as PubMed), patient data or health records. With such a huge amount of reports, evaluating their impact has long seized to be a trivial task. In this context, this paper intends to introduce a non-scientific factor that represents an important element in the effort of gaining acceptance of claims. Thus, we postulate that the confidence the author is expressing in his work plays an important role in shaping the first impression that influences the reader's perception of the paper. The results discussed in this paper are based on a series of experiments ran over data from the Open Archives Initiative (OAI) corpus that provides interoperability standards in order to facilitate the effectiveness dissemination of the content. This method can be useful to the direct beneficiaries (authors, who are engaged in medical or academic research), but, also, researchers in the fields of BioNLP and NLP, etc.

Keywords: Biomedical libraries; author's confidence; writing styles; text analysis

1. Summary

The interest for biomedical digital libraries, along with the continuous development of various qualitative and quantitative text analyses tools, made language technologies a natural choice for analyzing the evolution of the scientific life. Mining biomedical literature to extract the science behind it, such as concepts, patterns or relations, is a very productive research area. However, extracting non-scientific information from biomedical data has recently also seen an increasing interest, with applications ranging from identifying speculative language, to retrieval of papers with a specific writing style, in an attempt to cope with different reader preferences.

This paper proposes a method to identify the degrees of confidence that an author has in its own writing. Experiments and results discussed in this paper are based complex system, ran over a set of data extracted from the Open Archives Initiative (OAI) corpus[1], consisting of over 10.000 papers extracted for the timeframe 2006 – 2017 for the *malaria* domain.

This survey is based on the legitimate question: *What elements betray the author's level of trust in his own scientific writing?*

The paper is structured as follows: Section 2 presents briefly relevant mining biomedical literature that reveals a large interest for identify the features that drive readers to choose a particular scientific article. Section 3 describes shortly the Open Archives Initiative (OAI) corpus of full-text academic articles between 2016-2017; Section 4 presents the architectural components to identify critical features for evaluating author's confidence. Section 5 describes a new system based on a

[1] https://www.openarchives.org/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0206.v1

Peer-reviewed version available at Data 2019, 4, 18; doi:10.3390/data4010018

linguistic analysis of scientific biomedical articles at the lexical, syntactic, and semantic levels, and the results are presented at the Section 6. The limitations of this methodology, based on three linguistic characteristics that we have considered sufficient at this stage for recognizing author's confidence, are presented at the section 7. A challenge for future work is to find reliable linguistic cues that generalize full confidence in the accuracy and integrity of the some author's work.

2. Background

Biomedical text mining (BioNLP) uses sophisticated predictive models to understand, identify and extract concepts from a large collection of scientific texts in medicine [7], biology, biophysics, chemistry, etc., in order to discover knowledge which can add value to biomedical research.
Therefore, a wide range of language resources was developed, including complex lexicons, thesauri and ontologies that cover the entire spectrum of clinical concepts. Keizer [3, 4] and Cornet [1] described a terminological and typology system to provide a uniform conceptual understanding.

Aside from mining knowledge, a new research direction tries to identify the factors that drive readers to choose one scientific article instead of another.

The retrieval of important literature represents a day-to-day activity for PhD students and scientific researchers, for both finding the latest breakthrough or for compiling a state-of-the-art for an area of interest. In [15], a set of stylometric features are used to develop an author search tool which allows finding paragraphs written by a specific author or in a specific writing style, since they directly relate the author's writing style to the readability of textual content. Hyland [9] analyzed 240 texts to verify if self-citation and exclusive first person pronouns influence paper acceptance in 8 disciplines.

An important research direction in the biomedical domain is the identification of hedges (speculative and tentative statements). If for most natural language applications hedging can be safely ignored, but for the biomedical domain it is essential to properly identify if a relation between a drug and a disease is a fact or just a speculation. Friedman et al. [6] discuss uncertainty in radiology reports and identified five levels of certainty. Other studies in the speculative aspect of biomedical text annotate speculations and identify them through simple substring matching [10], using machine learning techniques with variants of the well-known "bag-of-words" approach [11] or as classification problems [17].

The inspiration for our research was the study in [20] investigating the relation between an individual's self-reported confidence and the influence they had within a freely interacting group. They concluded that the influence of an individual within a group was directly dependent on his or her confidence level.

In this context, we hypothesized that a confident scientific paper will rather be selected, either for reading or for approval in various scientific journals, than a similar paper, but written in a less confident manner. Therefore, we developed an instrument for identifying an author's confidence, based on his or her writing style and other linguistic clues, such as passive vs. active voice, first vs. third person, etc.

3. Data set

In order to identify author confidence, we collected a set of 10.000 documents belonging to the Open Archives Initiative (OAI) corpus, which contains articles from 2006 to 2017. OAI develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open access and institutional repository movements. Over time, OAI has established itself as promoter of broad access to digital resources for e-Scholarship, e-Learning, and e-Science.

The collection contains several XML files, each with around 25 scientific articles, selected from the OAI among those which to contain the term "malaria" in either the title or the abstract, and belong to the specified timeframe. The reason for selecting a specific disease was that we expect articles to be comparable with regard to the medical terms they use.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0206.v1

Peer-reviewed version available at *Data* **2019**, *4*, 18; doi:10.3390/data4010018

The first step in our processing involved a pre-processing of the XML files in order to split each article in a separate file, which is then feed to the author confidence detection system.

An excerpt of the structure of the XML files for each article is presented in Figure 1. Each article is divided into its composing sections, enclosed into the `<sec> tag`. Although the structure of each article is different, according to the specific requirements of the publishing journal, there are some common sections appearing in most scientific writings: abstract, introduction, methods, results, conclusions. Each section contains a <title> tag and a set of paragraphs (`<p> tags`). Due to space restrictions, only the introduction section and part of the results section is presented in Figure 1.


```xml
<sec>
  <title>Introduction</title>
  <p>Mammography and sonography are the standard imaging techniques for detection and evaluation of breast disease [<xref ref-type="bibr" rid="B1">1</xref>]. Mammography is the most established screening modality [<xref ref-type="bibr" rid="B2">2</xref>]. Especially in young women and women with dense breasts, sonography appears superior to mammography, and differentiation between solid tumours and cysts is easier. Sensitivity and specificity of sonography or mammography are higher if sonography and mammography are combined [<xref ref-type="bibr" rid="B3">3</xref>].</p>
  <p>It is generally accepted that MR mammography is the most sensitive technique for diagnosis of breast cancer, whereas the reported specificity of MR mammography varies [<xref ref-type="bibr" rid="B4">4</xref>,<xref ref-type="bibr" rid="B5">5</xref>,<xref ref-type="bibr" rid="B6">6</xref>,<xref ref-type="bibr" rid="B7">7</xref>,<xref ref-type="bibr" rid="B8">8</xref>,<xref ref-type="bibr" rid="B9">9</xref>,<xref ref-type="bibr" rid="B10">10</xref>,<xref ref-type="bibr" rid="B11">11</xref>,<xref ref-type="bibr" rid="B12">12</xref>]. In those studies, MR mammography was performed and evaluated by highly specialized radiologists in a research setting. It was therefore the purpose of the present prospective study to compare the validity of MR mammography with mammography and sonography in clinical routine practice. Findings for the three diagnostic methods documented on routine reports that were available to the surgeon preoperatively formed the basis of this comparison. Special emphasis was placed on the identification of multifocal and multicentric invasive disease.</p>
</sec>
<sec sec-type="methods">
  <title>Patients and methods</title>
  <sec>
  <sec>
  <sec>
</sec>
<sec>
  <title>Results</title>
  <p>All patients underwent breast surgery and all abnormal lesions identified by mammography, sonography or MR mammography were surgically removed. A total of 477 breast lesions were examined histologically, revealing the presence of 185 invasive cancers, 38 carcinomata <italic>in situ</italic> and 254 benign lesions (fibroadenoma, papilloma, intraductal or adenoid ductal hyperplasia, cystic mastopathia). There were four patients with malignant lesions in both breasts. In 42 patients multifocal tumours and in nine patients multicentric tumors were found on histological examination. Among the 185 invasive lesions, 178 were primary cancers, five were recurrences, one was metastatic and one was an angiosarcoma. The majority of invasive breast cancers were staged as pT1c (44%). Six per cent of tumors were detected in stage pT1a, 18% in stage pT1b, 25% in stage pT2, 3% in stage pT3 and 4% in stage pT4. The distribution of histopathological tumour types is shown in Table <xref ref-type="table" rid="T1">1</xref>. The mean age of patients was 58 years (range 19-85 years).</p>
  <p>The sensitivity of MR mammography was significantly higher than those of mammography and sonography (<italic>P</italic> &lt; 0.005 and <italic>P</italic> &lt; 0.05; Table <xref ref-type="table" rid="T2">2</xref>). The specificity of sonography was significantly higher than those of mammography and MR mammography (<italic>P</italic> &lt; 0.05 and <italic>P</italic> &lt; 0.005; Table <xref ref-type="table" rid="T2">2</xref>). The negative predictive values for sonography and MR mammography were significantly higher than that of mammography (<italic>P</italic> &lt; 0.05 and <italic>P</italic> &lt; 0.005; Table <xref ref-type="table" rid="T2">2</xref>). With regard to accuracy, no significant difference between the three modalities was found (Table <xref ref-type="table" rid="T2">2</xref>). Combining of all three diagnostic methods yielded the best results for detection of cancer (<italic>P</italic> &lt; 0.005; Table <xref ref-type="table" rid="T3">3</xref>). The sensitivity and negative predictive value for the combination of mammography and MR mammography, and the combination of sonography and MR mammography were significantly higher than those for the combination of mammograpy and sonography (<italic>P</italic> &lt; 0.05; Table <xref ref-type="table" rid="T3">3</xref>). The highest result for accuracy was seen for a combination of all three methods (<italic>P</italic> &lt; 0.05; Table <xref ref-type="table" rid="T3">3</xref>).</p>
  <p>Mammography was false-negative in 30 out of 184 invasive cancers, sonography was false-negative in 20 out of 185 cancers, and 10 out of 185 invasive cancers were missed by MR mammography. The majority of false-negative findings was found in stage1 disease, ductal carcinoma and grade 3 tumors (Table <xref ref-type="table" rid="T4">4</xref>). Of 10 invasive cancers missed by MR mammography, eight were found by mammography and sonography. By all three techniques, one invasive ductal carcinoma (pT1b) was misinterpreted as fibroadenoma. In another patient, a microinvasive lobular carcinoma of 5 mm diameter was not detected with mammography and MR mammography, whereas sonography detected a solid, benign tumour. MR mammography identified 10 invasive cancers (5.2%) that were missed by mammography and sonography, whereas one invasive cancer was found by mammography alone. By sonography alone, not a single case of invasive disease was detected when MR mammography or mammography were nonsuspected.</p>
  <p>The highest detection rate for multifocal invasive disease was seen with MR mammography, which identified 28 out of 42 (66.7%) histologically confirmed multifocal invasive cancers, whereas mammography and sonography both identified 11 (26.2%) of these cancers (<italic>P</italic> &lt; 0.05). The combination of all three diagnostic methods leads
```


Figure 1. An example of XML file extracted from the OAI corpus

4. Architecture

While the study of the connection between discourse patterns and personal identification of an author is decades old, the study of these patterns, using language technologies, is relatively recent. In the more recent tradition, we frame author's confidence prediction from a text as an important problem for the natural language processing domain. Confidence [13] is generally described as a state of being certain either that a hypothesis or prediction is correct or that a chosen course of action is the best or most effective. Different approaches consider the confidence in terms of "appropriateness" or "trustworthiness" [23], or correlate it to uncertainty. In [21] the authors describe a function theory, called Dempster-Shafer (D-S), for evaluating the confidence of an argumentation. In [22], a trust case framework is used to check the argumentation used to demonstrate the compliance with specific standards.

In the context of this study, a structured argumentation, although it plays an important role in the communication, is not enough. Automatically discovering if an author is confident or not in his argumentation is a challenging task, which involves finding author's sentiments, features to determine his writing style, as well as information about his mastering of the scientific field.

The architecture of our proposed system is presented in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0206.v1

Peer-reviewed version available at Data 2019, 4, 18; doi:10.3390/data4010018

In order to determine the confidence of an author in his work, we propose a system composed of three main modules: a preprocessing step, a parser, and a voting procedure. After extracting each article in a separate XML file, a preprocessing step extracts only the text, deleting all tags. Only two sections were analyzed for each article, the *Results and Conclusion* one, since we found in a previous study that in these sections authors are more likely to present their work in a confident or reluctant manner (see Appendix A).

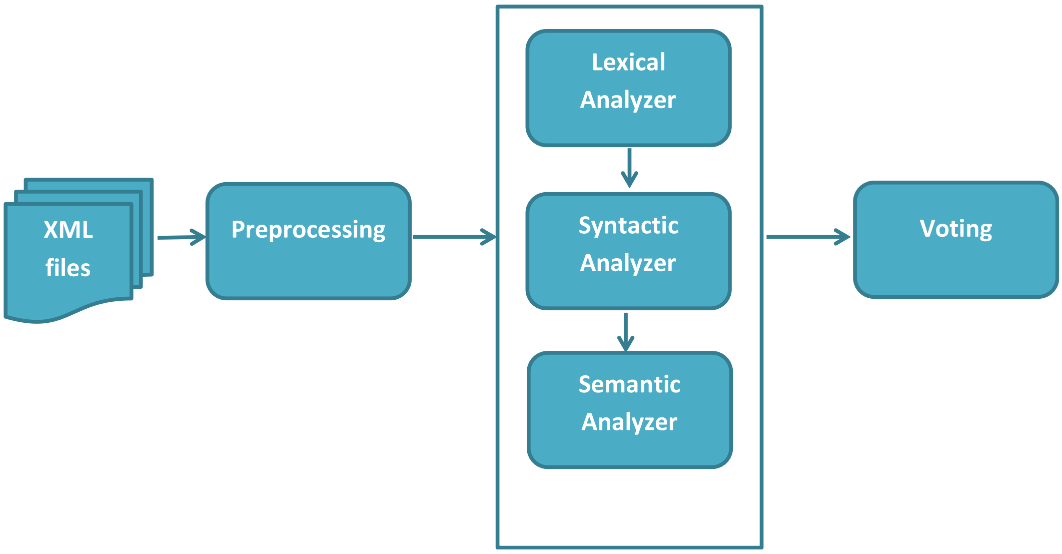

Figure 2. Architecture of our author confidence system

The raw text for the two sections is then cleaned to avoid sending unrecognized characters to the parser. The parser consists of three modules: a lexical, a syntactic and a semantic analyzer. The last step is the voting procedure, which takes the scores from the three previous analyzers and merges them using various thresholds, deciding if the text is written in a confident manner or not. The next section describes in more details the three analyzers.

5. System description

Our system is based on a linguistic analysis of scientific biomedical articles, by exploring various lexical, syntactic and semantic features. After the preprocessing step, the raw texts are feed to a parser with three modules, in a pipeline.

The first module is the lexical analyzer (see Figure 3), which tokenizes the text to identify each word. From this step, the sentence length can be obtained.

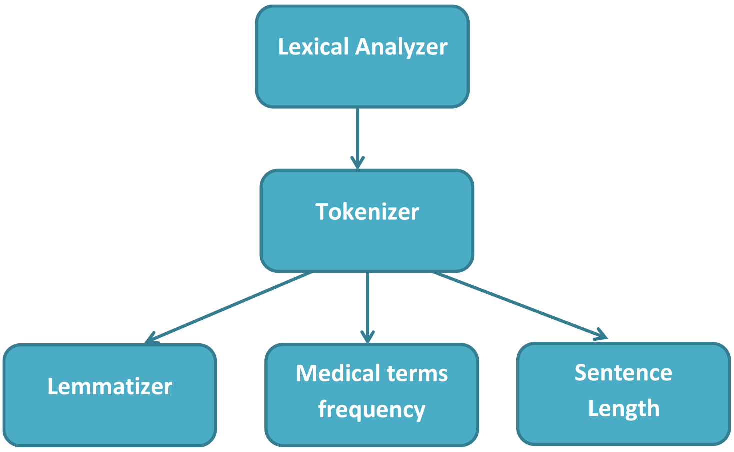

Figure 3. The lexical analyzer module

We analyzed this feature since we noticed that sentences which are too long tend to be more difficult to follow. After this step, a lemmatiser identifies the dictionary form of the words. This is

Peer-reviewed version available at Data 2019, 4, 18; doi:10.3390/data4010018

useful in order to count frequencies more accurately. Thus, unique unigrams and bigrams frequencies are computed and normalized by the length of the document and the number of tokens within. Functional words are removed, and the number of medical terms in each document is computed. Although a specialized language needs to be used to prove mastery of the domain, if the number of specialized words is too high in a document, when compared with words from the common vocabulary, the reading and understanding of the article has to suffer.

The second module is the syntactic analyzer, presented in figure 4. The part-of-speech (POS) tagging is performed using RACAI POS tagger[2]. Once parts of speeches have been identified, we extract two features: (1) the use of passive or active voice and (2) preference for using first or third person for both verbs and pronouns. We considered that the voice of the scientific articles is relevant since, in the argumentation theory, active voice is preferred and considered to indicate more commitment. The passive voice, on the contrary, indicates a certain distance from what is being presented.

For instance, the sentence:

"It has been showed that confident authors express themselves in active voice."

poses the accent on someone else, the one who made the statement, and establish a certain distance.

On the contrary, the active version of this sentence shows more commitment and agreement:

"Research has showed that confident authors express themselves in active voice."

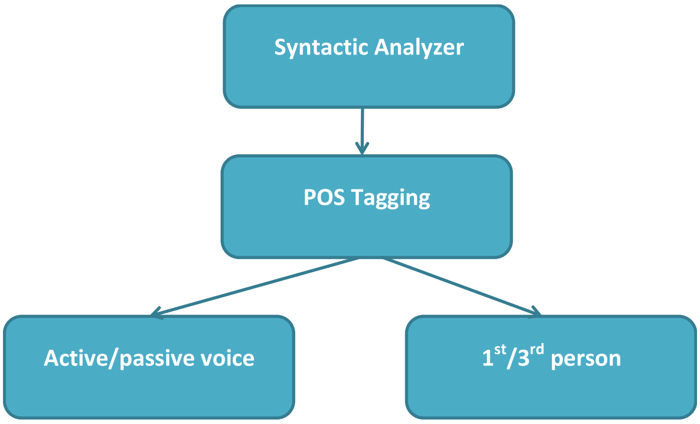

Figure 4. The syntactic analyzer module

Another relevant information is the inflection of the verbs and pronouns with regards to the number. Writing in the first, second, or third person is referred to as the author's point of view. The common tendency is to personalize the text of blogs, journal or books by writing in the first person ("I" and "we"). However, this tactic is not common in academic writings.

In science and mathematics, the first person is rarely used, being considered to move the focus of the statement from the research to the author. For medical text, it is in general acceptable to use the first person point of view in abstracts, introductions, discussions, and conclusions, in some journals to refer to the group of researchers that were part of the study. The third person point of view is used for writing methods and results sections. Adhering to this common practice shows knowledge of the usual norm, being in the same time a note of rigorousity and thus confidence.

The point of view of the third person is generally used in scientific papers, in different forms. Indefinite pronouns are used to refer back to the subject, while avoiding to have masculine or feminine terminology.

The following sentence uses the indefinite pronoun:

[2] http://www.racai.ro/en/tools/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0206.v1

Peer-reviewed version available at Data 2019, 4, 18; doi:10.3390/data4010018

`An author must ensure that he has used the proper person in his writing.`

An example of masculine and feminine terminology, which should be avoided, considered a factor of distraction if repeated, is:

`An author must ensure that he or she has used the proper person in his or her writing.`

The third and last module, named semantic analyzer, performs two types of analyses: sentiment identification and author profiling (see figure 5). The POS-tagged corpus of articles is filtered to identify the overall sentiment of each paper using Stanford Sentiment Analysis tool[3]. Their deep learning model builds up a representation of a whole sentence based on its grammatical structure. It computes the sentiment based on how words compose the meaning of longer phrases, using a Recurrent Neural Network. After analyzing each sentence individually, a score for the entire document is given.

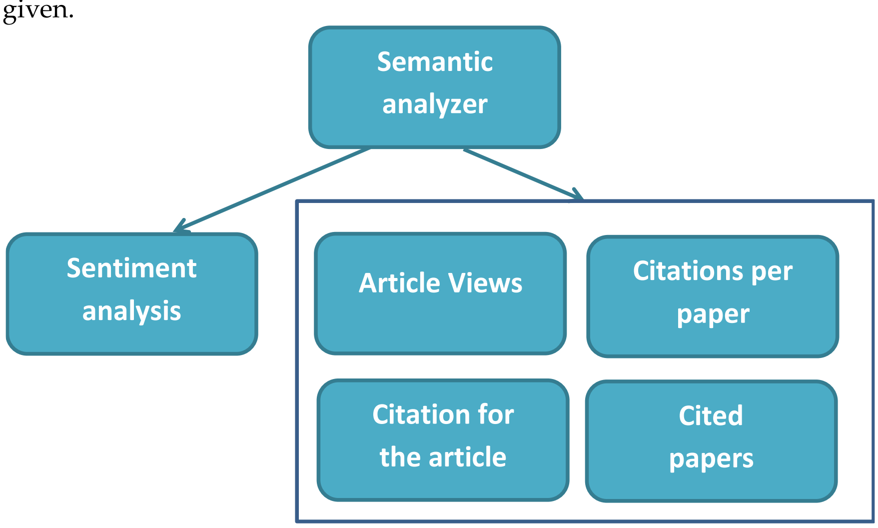

Figure 5. The semantic analyzer module

The collection of articles came with their own metadata (see Appendix B), from which we extracted information for the author's profile, i.e. name of the author, name of the journal, keywords, etc. In order to identify the importance of a paper for its domain, we looked over the internet to find the author's notoriety and investigated his previous publications by taking into account the number of citations per paper and the number of article views. Additionally, we considered the number of times the article was cited and the total number of cited reference papers for each given article.

Each of the three main analyzers (lexical, syntactic and semantic) returned a score for each article, and the final step involved the concatenation of the intermediate scores, with specific weights, in order to obtain the final result, which is a good predictor of whether a certain author has written his paper in a confident tone or not. The weights of each module are empirically identified, using information from the corpus, but also from various good practice guides on how to write a scientific article.

6. Results

This section presents the results obtained for three features (sentiment analysis, average number of words per sentence and frequency of medical terms) in evaluating an author's confidence (Figures 6, 7, and 8 respectively). We observe that through sentiment analysis and medical terms frequency we obtain distinctive results, suggesting that the choice of words of confident authors reflect positive sentiments, and medical terms frequency is in tandem with the first feature. The feature based on the average words per sentence had an irregular behavior. It is normal because the

[3] https://nlp.stanford.edu/sentiment/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0206.v1

Peer-reviewed version available at Data 2019, 4, 18; doi:10.3390/data4010018

performance of a good argumentation, in both spoken and written form, contains no unnecessary words.

6.1. Sentiment Analysis

The computational treatment of sentiments, subjectivity and opinions has recently attracted a great deal of attention, in part because of its potential applications. The sentiment analysis has proven useful for editorial sites, and companies to create summaries of people's experiences and opinions that consist of subjective expressions extracted from reviews or even just a review's polarity - positive or negative.

Identification of author's confidence poses a significant challenge to data-driven methods, resisting traditional techniques. In the present study, we used Sentiment Analysis in order to identify the author's level of confidence. In Figure 6, we show the results obtained after running the Sentiment Analysis tool.

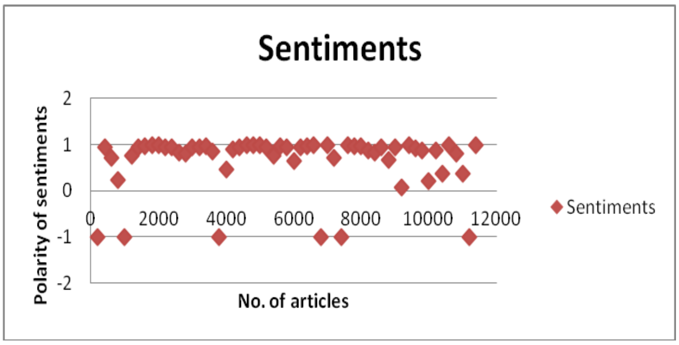

Figure 6. Visualization of the sentiment analysis

Our results indicate that most of the papers have positive (towards 1.5) sentiments, and that confidence is directly linked to positive expression of sentiment.

6.2. Average words per sentence

When writing a scientific paper, the first quality, with precedence over all others, is clarity. According to Oxford Academy[4], it is highly recommended to use up to 15 words in a sentence, and if an author chooses to use too many word in a sentence, it reveals a low degree of confidence while writing the work in question. This analysis is supported by our findings, the article marked as having a confident author having the average sentence length in the rage of 15-20 words.

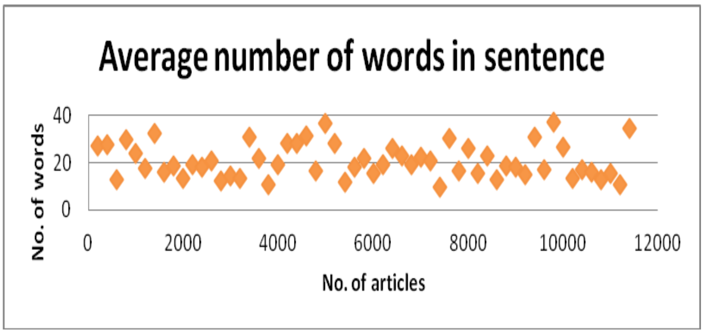

Figure 7. Average number of words per sentence

6.3. Medical frequency terms

[4]https://www.ox.ac.uk/sites/files/oxford/field/field_document/Tutorial%20essays%20for%20science%20subjects.pdf

To demonstrate that an author is self-confident, it is essential to use the appropriate terms (in our case, the medical terms), to avoid jargon, because it is the secret language of the scientific field. It excludes the intelligent, otherwise well-informed, reader, and speaks only to the initiated. The analysis of our corpus showed that the articles marked with non-confidence had either below 25% of medical terminology, or above 40%.

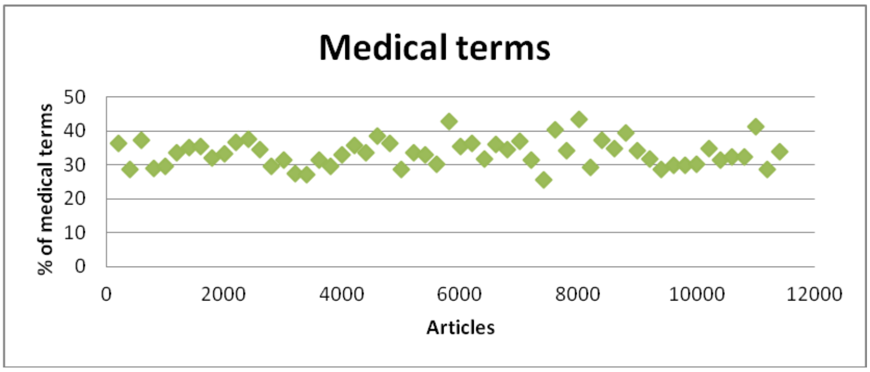

Figure 8. Medical frequency terms chart

In this study, we have shown that it is possible to automatic identify the level of confidence an author had when writing a scientific paper.

7. Discussions

In this paper we have presented a method to extract non-scientific information form biomedical papers, more specifically the confidence of an author about his work. Given this purpose, we explore linguistic features that are predictive of author's level of trust in his own scientific writing. While our focus is on a single type of disease ("malaria"), we choose a method that generalizes to other diseases, revealing the similarity present in other medical interactions.

We have studied the relation between lexical analysis (frequencies of medical words, sentence length), syntactic features (POS tagging, voice and person of verbs and pronouns) and semantic features (sentiment analysis, author profiling) in order to automatically predict the author's confidence.

To improve the performance of our system, we intend to enrich the gold annotated corpus with articles for different diseases and use additional machine learning techniques.

To further test our belief that author confidence influences the acceptance of papers in peer-reviewed journals, we intend to extend the study by analyzing reviews from journals with open review process.

Supplementary Materials: The following are available online at https://profs.info.uaic.ro/~daniela.gifu/LR/PubMed%20Corpus_Malaria/, Figure 1: title, Figure 2: title, Figure 3: title, Figure 4: title, Figure 5: title, Figure 6: title, Figure 7: title, Figure 8: title.

Author Contributions: Conceptualization, M.O. and D.G.; methodology, M.O. and D.T.; software, D.T; validation, M.O and D.G.; formal analysis, M.O. and D.T.; investigation, D.G.; resources, M.O.; data curation, M.O and D.T.; writing-original draft preparation, M.O. and D.G.; editing and revision, M.O. and D.T.; supervision, D.G. and D.T.; project administration, D.G. and D.T.; funding acquisition, D.G. and D.T.

Funding: This research was funded by This project was partially supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818 / 73PCCDI (ReTeRom), within PNCDI III and by the README project "Interactive and Innovative application for evaluating the readability of texts in Romanian Language and for improving users' writing styles", contract no. 114/15.09.2017, MySMIS 2014 code 119286.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A - An example of chapters analyzed, *Results and Conclusion*, in XML format

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0206.v1

Peer-reviewed version available at Data 2019, 4, 18; doi:10.3390/data4010018

<title>Conclusion</title>

<p>This paper studied multiple affiliations of authors in research publications. Results for three scientific fields (biology, chemistry and engineering) and three countries (Germany, Japan and the UK) showed that multiple affiliations are widespread and have increased in all fields and countries during the period 2008–2014.</p>

<p>We found that multiple affiliations reflect the dynamics of the research sector in specific countries and proposed a classification of the cross-sector and international dimension of author affiliations. To summarise, we find three types of multiple affiliations that can be classified as (A) a highly internationalised, HEI cantered affiliation distribution as represented by researchers in the UK, (B) a balanced affiliation distribution as seen in Germany, and (C) a domestic, cross-sector affiliation distribution as seen in Japan. These results suggest that cross-sector affiliations are highest in countries and fields with a large non-university research sector, while cross-country affiliations are highest in countries with an international research base. An analysis of other countries may find additional types. However, the occurrence of low cross-sector affiliations paired with low internationalisation, that is, where academic authors are primarily affiliated with other domestic universities, may be limited by academic employment contracts which generally still limit such arrangements.</p>

<p>These observed differences have consequences for the types of networking that can be achieved through multiple affiliations in different countries. For example, international affiliations may help to preserve links to 'frontline' research institutions, while cross-sector affiliations may be more conducive to knowledge transfer and mobility between sectors (ESF <xref ref-type="bibr" rid="CR5">2013</xref>). Our results did, however, show that most multiple affiliations of academics are with other universities or with PROs, including in the cases of Japan and Germany. The role of multiple affiliations as a facilitator for knowledge transfer between distinct sectors (ESF <xref ref-type="bibr" rid="CR5">2013</xref>) may therefore be rather limited.</p>

<title>Results</title>

<p>Table <xref rid="Tab1" ref-type="table">1</xref> shows the total number of authors reported on the selected publications by country and field, as well as the number and proportion of authors that report more than one institutional address. Of the more than 118,000 authors in the sample, 7.2% have more than one institution attached, with some differences across countries and subject areas.<xref ref-type="fn" rid="Fn5">5</xref> The proportion of authors with multiple institutional addresses is highest with more than 9% of authors in biology and chemistry in the case of Germany, and biology in the case of the UK. This already suggests some country and subject-specific differences regarding the extent of multiple affiliations.</p>

Appendix B - An example of metadata for a scientific article on malaria issue, in XML format

doi:10.20944/preprints201811.0206.v1

Peer-reviewed version available at *Data* **2019**, *4*, 18; doi:10.3390/data4010018


```xml
<journal-meta>
      <journal-id journal-id-type="nlm-ta">Curr Hematol Malig Rep</journal-id>
      <journal-id journal-id-type="iso-abbrev">Curr Hematol Malig Rep</journal-id>
      <journal-title-group>
        <journal-title>Current Hematologic Malignancy Reports</journal-title>
      </journal-title-group>
      <issn pub-type="ppub">1558-8211</issn>
      <publisher>
        <publisher-name>Springer US</publisher-name>
        <publisher-loc>New York</publisher-loc>
      </publisher>
    </journal-meta>
    <article-meta>
      <article-categories>
        <subj-group subj-group-type="heading">
          <subject>Stem Cell Transplantation (R Maziarz, Section Editor)</subject>
        </subj-group>
      </article-categories>
      <title-group>
        <article-title>10 Years of Preparedness by the Radiation Injury Treatment Network</article-title>
      </title-group>
      <contrib-group>
        <contrib contrib-type="author" corresp="yes">
          <name>
            <surname>Case</surname>
            <given-names>Cullen</given-names>
            <suffix>Jr.</suffix>
          </name>
          <address>
          </address>
          <xref ref-type="aff" rid="Aff1"/>
        </contrib>
      </contrib-group>
      <pub-date pub-type="epub">
        <year>2017</year>
      </pub-date>
      <volume>12</volume>
      <issue>1</issue>
      <fpage>39</fpage>
      <lpage>43</lpage>
<bold>Open Access</bold> This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses,
        </license>
      </permissions>
      <abstract id="Abs1">
      <kwd-group xml:lang="en">
        <title>Keywords</title>
        <kwd>Radiological</kwd>
        <kwd>Emergency</kwd>
      </kwd-group>
      <custom-meta-group>
        <custom-meta>
          <meta-name>issue-copyright-statement</meta-name>
          <meta-value>© Springer Science+Business Media New York 2017</meta-value>
        </custom-meta>
      </custom-meta-group>
    </article-meta>
```


References

1. Cornet R., de Keizer N.F., Abu-Hanna A. *A framework for characterizing terminological systems.* In: *Methods Inf Med.*, 2006, vol. 45, pp. 253-266.
2. Dashevskiy M., Luo Z. *Predictions with Confidence in Applications*. In: Perner P. (eds) *Machine Learning and Data Mining in Pattern Recognition. MLDM*, 2009, pp. 775-786.
3. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonl JHM. *Understanding terminological systems I: terminology and typology.* In: *Methods Inf Med*., 2000, vol. 39, pp. 16-21.
4. de Keizer, NF, Abu-Hanna, A. *Understanding terminological systems II: terminology and typology*. In: *Methods Inf Med*., 2000, vol. 39, pp. 22-29.
5. Derntl, M. *Basics of research paper writing and publishing*, In: *J. Technology Enhanced Learning*, 2014, Vol. 6, No. 2, http://dbis.rwth-aachen.de/~derntl/papers/misc/paperwriting.pdf
6. Friedman C., Alderson P, Austin J, Cimino JJ, Johnson SB. *A general natural-language text processor for clinical radiology. J Am Med Info Assoc*., 1994, pp. 1: 161-174.
7. Gifu, D. *Malaria Detection System*. In: *Proceedings of the International Conference on Mathematical Foundations of Informatics*, MFOI-2017, Cojocaru, S. and Gaindric, C., Druguș, I. (eds.), Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, Chișinău, 2017, pp.74-78.
8. Hangyo, M., Kawahara, D., Kurohashi, S. Japanese Zero. *Reference Resolution Considering Exophora and Author/ReaderMentions*, http://aclweb.org/anthology/D/D13/D13-1095.pdf.
9. Hyland, K. *Humble servants of the discipline? Self-mention in research articles in English for Specific Purposes*, Volume 20, Issue 3, 2001, Pages 207-226.
10. Light M., Qiu, X.Y., Srinivasan P. *The language of bioscience: facts, speculations, and statements in between.* In: *BioLINK 2004: Linking Biological Literature, Ontologies and Databases*, 2004, pp. 17-24
11. Medlock, B., Briscoe, T. (2007). *Weakly supervised learning for hedge classification in scientific literature*. In: *Proceedings of 45th Meeting of the Association for Computational Linguistics*, 2007, pp. 992-999.
12. Nguyen, D., Smith, N.A., Rose, C.P. (2011). *Author Age Prediction from Text using Linear Regression*, http://aclweb.org/anthology/W/W11/W11-1515.pdf.
13. Partridge, D., Bailey, T.C., Everson, R.M., Fieldsend, J.E., Hernandez, A., Krzanowski, W.J., Schetinin, V. *Classification with Confidence for Critical Systems.* In: *Developments in Risk-based Approaches to Safety*, 2007, pp. 231–239.

Peer-reviewed version available at *Data* **2019**, *4*, 18; doi:10.3390/data4010018

14. Qian, T., Liu, B. *Identifying Multiple Userids of the Same Author*, 2013, http://aclweb.org/anthology/D/D13/D13-1113.pdf.
15. Rexha, A., Kröll, M., Ziak, H., & Kern, R. *Extending Scientific Literature Search by Including the Author's Writing Style*. In: BIR@ ECIR, 2017, pp. 93-100.
16. Sim, Y., Routledge, B.R., Smith, N.A. *A Utility Model of Authors in the Scientific Community*, 2015, http://aclweb.org/anthology/D/D15/D15-1175.pdf.
17. Szarvas, G. *Hedge classification in biomedical texts with a weakly supervised selection of keywords.* In: *Proceedings of 46th Meeting of the Association for Computational Linguistics.* 2008, pp. 281-289.
18. Thompson P., Venturi G., McNaught J., Montemagni S., Ananiadou S. *Categorising modality in biomedical texts.* In: *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, 2008, pp. 27-34.
19. Wilbur W.J., Rzhetsky A., Shatkay H. *New directions in biomedical text annotations: definitions, guidelines and corpus construction.* In: *BMC Bioinformatics*, 2006, pp. 7: 356.
20. Zarnoth, P., Sniezek, J. *The Social Influence of Confidence in Group Decision Making*. In: *Journal of Experimental Social Psychology*, Volume 33, Issue 4, July 1997, pp. 345-366.
21. Richard Hawkins, Tim Kelly, John Knight, and Patrick Graydon. A new approach to creating clear safety arguments. In Advances in systems safety, pages 3–23. Springer, 2011.
22. Wang, R., Guiochet, J., Motet, G., Schön, W. *D-S Theory for Argument Confidence Assessment*. 4th International Conference on Belief Functions (BELIEF 2016), Sep 2016, Prague, Czech Republic. 2016, pp.190-200.
23. Cyra, L. and Gorski, J. *Supporting compliance with security standards by trust case templates*. In Dependability of Computer Systems, 2007. DepCoS-RELCOMEX'07. 2nd International Conference on, pp. 91–98. IEEE.