data

MDPI

Article

Towards Identifying Author Confidence in Biomedical Articles

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

[1] Institute of Computer Science, Romanian Academy-Iasi branch, 700481 Iasi, Romania; mihaela.onofrei@iit.academiaromana-is.ro
[2] Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, 700483 Iași, Romania; dtrandabat@info.uaic.ro
[3] Cognos Business Consulting S.R.L., 7, Iuliu Maniu Blvd, 061072 Bucharest, Romania
* Correspondence: daniela.gifu73@gmail.com; Tel.: +40-742-050-673

Received: 6 November 2018; Accepted: 17 January 2019; Published: 21 January 2019

Abstract: In an era where the volume of medical literature is increasing daily, researchers in the biomedical and clinical areas have joined efforts with language engineers to analyze the 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 stopped being a trivial task. In this context, this paper intended to introduce a non-scientific factor that represents an important element in gaining acceptance of claims. We postulated that the confidence that an author has in expressing their 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 were based on a series of experiments that were ran using data from the open archives initiative (OAI) corpus, which provides interoperability standards to facilitate effective dissemination of the content. This method may be useful to the direct beneficiaries (i.e., authors, who are engaged in medical or academic research), but also, to the researchers in the fields of biomedical text mining (BioNLP) and NLP, etc.

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

1. Introduction

The interest in biomedical digital libraries, along with the continuous development of various qualitative and quantitative text analysis tools, has made language technologies the natural choice to analyze the evolution of scientific life. Mining biomedical literature to extract the science behind it, such as concepts, patterns, or relations, is a very productive research area. The extraction of non-scientific information from biomedical data has recently seen an increase in interest, with applications ranging from the identification of speculative language, to the 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 degree of confidence that an author has in their own writing. The experiments and results discussed in this paper are based on a complex system, that was run using a set of data extracted from the open archives initiative (OAI) corpus (https://www.openarchives.org/), which consists of over 12,000 papers extracted for the timeframe 2006–2017 under the malaria domain.

This survey was based on the legitimate question: What elements reveal the author's level of trust in their own scientific writing?

The paper is structured as follows: Section 2 briefly presents the relevant articles regarding the mining of biomedical literature, which reveals a wide interest in identifying features that drive readers to choose a particular scientific article. Section 3 shortly describes the open archives initiative (OAI)

corpus of full-text academic articles. Section 4 presents the architectural components used to identify the critical features for evaluating authors' confidence. Section 5 describes a new system based on the linguistic analysis of scientific biomedical articles at the lexical, syntactic, and semantic level, and the results are presented in Section 6. The limitations of this methodology, which focused on three linguistic characteristics for recognizing authors' confidence that were closely analyzed at this stage, are presented in Section 7. A challenge for future work is to find reliable linguistic cues that generalize full confidence in the accuracy and integrity of the 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 the fields of medicine [1], biology, biophysics, chemistry, etc., to discover knowledge which can add value to biomedical research [2].

Therefore, a wide range of language resources were developed, including complex lexicons, thesauri, and ontologies that covered the entire spectrum of clinical concepts. Keizer [3,4] and Cornet [5] described a terminological and typology system to provide a uniform conceptual understanding.

Aside from mining knowledge, part of this 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, both for finding the latest breakthrough or for compiling a state-of-the-art for an area of interest. In [6], a set of stylometric features were used to develop an author search tool which allowed the finding of paragraphs written by a specific author or in a specific writing style, since they directly related the author's writing style to the readability of textual content [7–9]. Hyland [10] analyzed 240 texts to verify if self-citation and exclusive first person pronouns influenced paper acceptance in eight disciplines.

An important research direction in the biomedical domain is the identification of hedges (i.e., speculative and tentative statements). For most natural language applications hedging can be safely ignored; however, for the biomedical domain it is essential to properly identify if a relation between a drug and a disease is a fact or just speculation. Friedman et al. [11] discuss uncertainty in radiology reports and they identified five levels of certainty. Other studies in the speculative aspect of biomedical text annotate speculations [12] and identify them through simple substring matching [13,14], using machine learning techniques with variants of the well-known "bag-of-words" approach [15] or as classification problems [16,17].

The inspiration for our research was the study in [18], which investigated the relation between an individual's self-reported confidence and the influence that 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 be selected, either for reading or for approval in various scientific journals, compared to a similar paper, 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 versus active voice, first versus third person, etc.

3. Data Set

To identify author confidence, we collected a set of about 12,000 documents belonging to the open archives initiative (OAI) corpus, which contains articles from 2006 to 2017, in English. 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 a promoter of broad access to digital resources for e-Scholarship, e-Learning, and e-Science.

The collection contained several XML files, each with around 25 scientific articles, selected from the OAI amongst articles which contained the term "malaria" in either the title or the abstract, and also belonged to the specified timeframe. The reason for selecting a specific disease was that we expected the articles to be comparable with regard to the medical terms that were used.

The first step in our processing involved a pre-processing of the XML files to split each article to a separate file, which was then fed 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 was divided into its composing sections, enclosed into the `<sec> tag`. Although the structure of each article was different, according to the specific requirements of the publishing journal, there were some common sections appearing in most scientific writings: Abstract, introduction, methods, results, conclusions. Each section contained a <title> tag and a set of paragraphs (`<p> tags`). Owing to space restrictions, only the introduction section and part of the results section are presented in Figure 1.

```
<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 open archives initiative (OAI) corpus.

4. Methodology

While the study of the connection between discourse patterns and the 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 framed the author's confidence prediction from a text as an important problem for the natural language processing domain. Confidence [19] is generally described as a state of being certain that either a hypothesis or prediction is correct or that a chosen course of action is the best or most effective. Different approaches consider confidence in terms of "appropriateness" or "trustworthiness" [20], or they correlate it to uncertainty. In [21], the authors described a function theory, called Dempster–Shafer (D-S), for evaluating the confidence of an argumentation. In [22], a trust case framework was used to check the argumentation used to demonstrate the compliance with

specific standards. The programming language used was Python, with its most useful package the NLTK (Natural Language ToolKit).

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 the author's sentiments, features to determine their writing style, as well as information about the author's mastering of the scientific field.

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

In order to determine the confidence of an author in their work, we proposed 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, the preprocessing step extracted only the text, whilst deleting all the tags. Only two sections were analyzed for each article, the Results and Conclusion sections, since we found in a previous study that in these sections, authors were more likely to present their work in a confident or reluctant manner (see Appendix A).

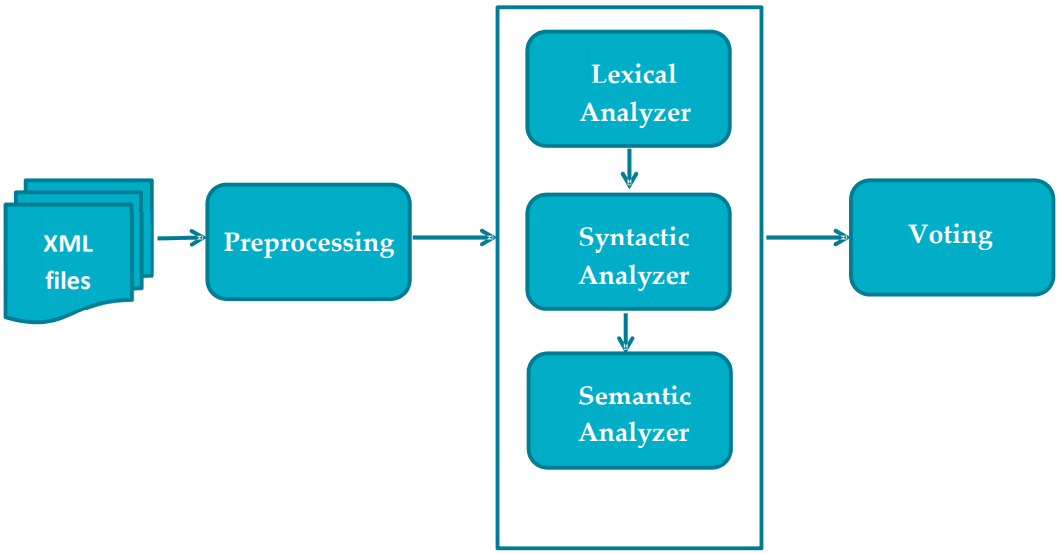

Figure 2. Architecture of our author confidence system.

The raw text of the two sections was then cleaned to avoid sending unrecognized characters to the parser. The parser consisted of three modules: a lexical, a syntactic, and a semantic analyzer. The last step was the voting procedure, which took the scores from the three previous analyzers and merged them. The weights of each feature were empirically determined, based on annotated examples, but also on the feature relevance. The system was initially tested with equal weights for all features, and the performance was recorded. Subsequently, the system was fine-tuned by running it with different values for the weights, and then saving the best performing version. Then, we used a threshold to decide if the text was written in a confident manner or not. The next section describes the three analyzers in more detail.

5. System Description

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

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

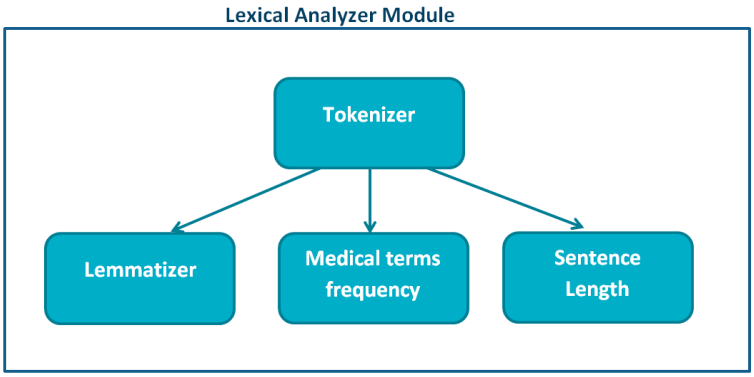

Figure 3. Description of the lexical analyzer.

We analyzed this feature, since we noticed that sentences which were too long tended to be more difficult to follow. After this step, a lemmatiser identified the dictionary form of the words. This was useful in order to count the frequencies more accurately. Thus, the frequencies of unique unigrams and bigrams were computed and normalized by the length of the document and the number of tokens within. Functional words were removed, and the number of medical terms in each document was computed. Although 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 a comparison is made with words from the common vocabulary, the reading and understanding of the article becomes difficult.

The second module was the syntactic analyzer, presented in Figure 4. The part-of-speech (POS) tagging was performed using a RACAI POS tagger (http://www.racai.ro/en/tools/) for English. We chose this tagger instead of the NLTK's POS functionality, since it facilitated the extraction of the verbal voice, a feature that we further used. Once parts of the speeches had been identified, we extracted two features: (1) The use of a passive or active voice, and (2) the preference for using the 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, the 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 shown that confident authors express themselves in an active voice."

focuses on someone else, i.e., the one who made the statement, and it establishes a certain distance.

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

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

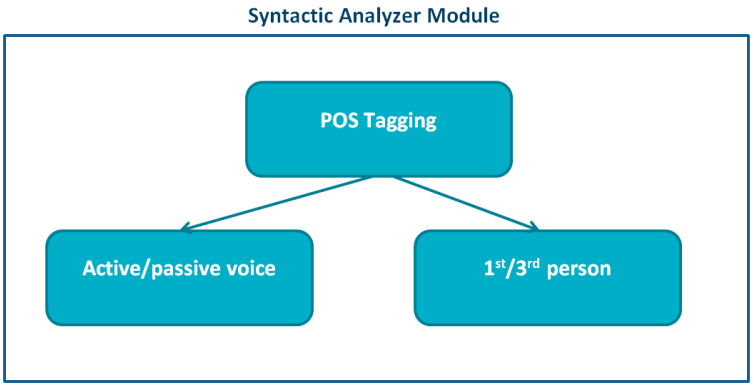

Figure 4. Description of the syntactic analyzer.

The other relevant information was 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 [23]. The common tendency is to personalize the text of blogs, journals, 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. In medical texts, in some journals, it is generally acceptable to use the first person point of view in the abstracts, introductions, discussions, and conclusions, to refer to the group of researchers that were part of the study. The third person point of view is used when writing the methods and results sections. Adhering to this common practice shows knowledge of the usual norm, as well as showing rigorousness, and thus confidence.

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

The following sentence uses the indefinite pronoun:

"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 the semantic analyzer, performed two types of analyses (see Figure 5): Sentiment identification and author profiling. The POS-tagged corpus of the articles was filtered to identify the overall sentiment of each paper using the Stanford sentiment analysis tool (https://nlp.stanford.edu/sentiment/). Their deep learning model builds up a representation of a whole sentence based on its grammatical structure. The Stanford sentiment analysis tool computes the sentiment based on how words compose the meaning of longer phrases, using a recurrent neural network. The sentiment is expressed as a polarity (i.e., the text tends to be positive or negative). After analyzing each sentence individually, a score for the entire document is given.

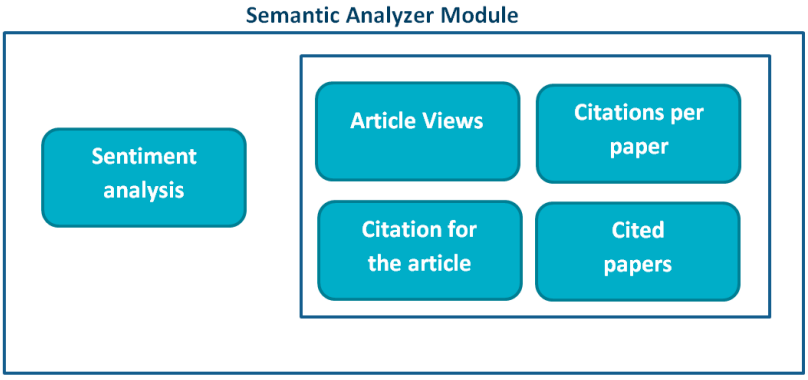

Figure 5. The semantic analyzer module.

The collection of articles came with its own metadata (see Appendix B), from which we extracted information about the author's profile, i.e., the name of the author, name of the journal, keywords, etc. In order to identify the importance of a paper in its domain, we checked the internet to find the author's notoriety and investigated the author's previous publications by considering 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, to obtain the final result, which was a good predictor of whether a certain author had written their paper in a

confident tone or not. The weights of each module were empirically identified, using information from the corpus, but also from various online good practice guides on how to write a scientific article.

For example, for the first module, the lexical module, the sentence length and the frequency of medical terms formed the score for the module. Thus, the weight for sentence length was 0.05 if the sentence was between 15 and 20 words, and 0 otherwise. Concerning the frequency of medical terms, if they appeared at less than 33% of all words in the article, the weight of these features was 0.75× normalized frequency, otherwise we used the weight 0.25. The linear combination of these two scores formed the overall score for the first lexical module. Similarly, weights were computed for the active vs. passive voice and the 1st vs. 3rd person, and their combination formed the score for the syntactic module. As for the third, the semantic module, a weight of 0.5× sentiment score was used for the sentiments identified, to which the author profile score was added. The latter score was computed as the sum of its different components: (a) 0.05× publication number/10 for authors with less than 10 publications, (b) 0.05; otherwise, 0.15× number of views/1000 for less than 1000 views, 0.05 otherwise, and (c) a citation score, was similarly computed. All the three scores (lexical, syntactic, and semantic scores) had equal weights in the total score.

6. Results

This section presents the results obtained for the three features (sentiment analysis, average number of words per sentence, and the frequency of medical terms) in evaluating an author's confidence (Figure 6, Figure 7, and Figure 8 respectively). We observed that through the sentiment analysis and medical terms frequency we obtained distinctive results, suggesting that the choice of words of confident authors reflected positive sentiments, and the medical terms frequency was in tandem with the first feature. The feature based on the average words per sentence had an irregular behavior. It was normal because the performance of a good argumentation, in both spoken and written form, contained 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. Sentiment analysis has proven useful for editorial sites. Companies create summaries of peoples' experiences and opinions are extracted from reviews based on a review's polarity, i.e., positive or negative.

Identification of the author's confidence poses a significant challenge to data-driven methods, resisting the traditional techniques. In the present study, we used sentiment analysis 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 indicated that most of the papers had positive (towards 1) sentiments, and that confidence was directly linked to the 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 the Oxford Academy (https://www.ox.ac.uk/sites/files/oxford/field/field_document/Tutorial%20essays%20for%20science%20subjects.pdf), it is highly recommended to use up to 15 words in a sentence, and if an author chooses to use too many words in a sentence, it reveals a low degree of confidence while writing the work in question. This analysis was supported by our findings, where an article that was marked as having a confident author would have an average sentence length in the range of 15–20 words.

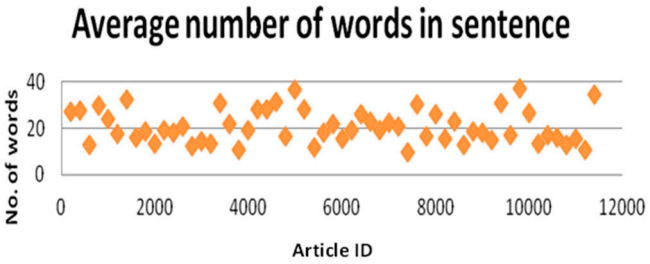

Figure 7. Average number of words per sentence.

6.3. Medical Frequency Terms

To demonstrate that an author is self-confident, it is essential to use the appropriate terms (in our case, the medical terms), and 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 statistical analysis of our corpus showed that the articles marked as showing non-confidence had either below 25% of medical terminology or above 40%.

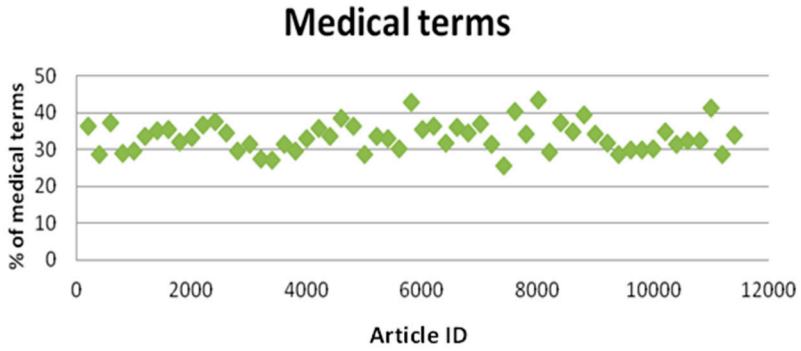

Figure 8. Medical frequency terms chart.

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

7. Discussion

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

We 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), to automatically predict the author's confidence. The weights

of each feature were empirically determined, based on annotated examples, but also on the feature's own relevance.

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

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 the reviews from journals with an open review process.

Author Contributions: Conceptualization, M.O.P. and D.T.; methodology, M.O.P. and D.G.; software, M.O.P.; validation, M.O.P. and D.G.; formal analysis, M.O.P. and D.T.; investigation, D.G.; resources, M.O.P.; data curation, M.O.P. and D.T.; writing-original draft preparation, M.O.P. and D.G.; editing and revision, D.G. 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 partially supported by a grant from 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

```
<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 summarize, we find three types of multiple affiliations
that can be classified as (A) a highly internationalized, 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 internationalization, 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

```
<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. Gifu, D. Malaria Detection System. In *Institute of Mathematics and Computer Science, Proceedings of the International Conference on Mathematical Foundations of Informatics (MFOI-2017), Chișinău, Moldova, 9–11 November 2017*; Cojocaru, S., Gaindric, C., Druguș, I., Eds.; Academy of Sciences of Moldova: Chișinău, Moldova, 2017; pp. 74–78.
2. Dashevskiy, M.; Luo, Z. Predictions with Confidence in Applications. In Proceedings of the Machine Learning and Data Mining in Pattern Recognition, Leipzig, Germany, 23–25 July 2009; pp. 775–786.
3. De Keizer, N.F.; Abu-Hanna, A.; Zwetsloot-Schonl, J.H.M. Understanding terminological systems I: Terminology and typology. *Method. Inf. Med.* **2000**, *39*, 16–21.
4. De Keizer, N.F.; Abu-Hanna, A. Understanding terminological systems II: Terminology and typology. *Method Inf. Med.* **2000**, *39*, 22–29.
5. Cornet, R.; de Keizer, N.F.; Abu-Hanna, A. A framework for characterizing terminological systems. *Method. Inf. Med.* **2006**, *45*, 253–266.
6. Rexha, A.; Kröll, M.; Ziak, H.; Kern, R. Extending Scientific Literature Search by Including the Author's Writing Style. In Proceedings of the BIR@ ECIR, Abeerden, UK, 8–12 April 2017; pp. 93–100.

7. Hangyo, M.; Kawahara, D.; Kurohashi, S. Japanese Zero. Reference Resolution Considering Exophora and Author/ReaderMentions. Available online: http://aclweb.org/anthology/D/D13/D13-1095.pdf (accessed on 20 January 2019).
8. Nguyen, D.; Smith, N.A.; Rose, C.P. Author Age Prediction from Text using Linear Regression. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland, OR, USA, 24 June 2011; Available online: http://aclweb.org/anthology/W/W11/W11-1515.pdf (accessed on 20 January 2019).
9. Qian, T.; Liu, B. Identifying Multiple Userids of the Same Author. In Proceedings of the Conference on Empirical Methods in Natural Language, Seattle, WA, USA, 18–21 October 2013; Available online: http://aclweb.org/anthology/D/D13/D13-1113.pdf (accessed on 20 January 2019).
10. Hyland, K. Humble servants of the discipline? Self-mention in research articles in English for Specific Purposes. *Engl. Specif. Purp.* **2001**, *20*, 207–226. [CrossRef]
11. Friedman, C.; Alderson, P.; Austin, J.; Cimino, J.J.; Johnson, S.B. A general natural-language text processor for clinical radiology. *J. Am. Med. Inf. Assoc.* **1994**, *1*, 161–174. [CrossRef]
12. Wilbur, W.J.; Rzhetsky, A.; Shatkay, H. New directions in biomedical text annotations: Definitions, guidelines and corpus construction. *BMC Bioinform.* **2006**, *7*, 356. [CrossRef] [PubMed]
13. Light, M.; Qiu, X.Y.; Srinivasan, P. The Language of Bioscience: Facts, Speculations, and Statements in Between. In Proceedings of the BioLINK 2004: Linking Biological Literature, Ontologies and Databases, Boston, MA, USA, 6 May 2004; pp. 17–34.
14. 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, Maarakech, Morocco, 26 May 2008; pp. 27–34.
15. Medlock, B.; Briscoe, T. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In Proceedings of the 45th Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 25–27 June 2007; pp. 992–999.
16. Sim, Y.; Routledge, B.R.; Smith, N.A. A Utility Model of Authors in the Scientific Community. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1510–1519. Available online: http://aclweb.org/anthology/D/D15/D15-1175.pdf (accessed on 20 January 2019).
17. Szarvas, G. Hedge Classification in Biomedical Texts with a Weakly Supervised Selection of Keywords. In Proceedings of the 46th Meeting of the Association for Computational Linguistics, Columbus, OH, USA, 16 June 2008; pp. 281–289.
18. Zarnoth, P.; Sniezek, J. The social influence of confidence in group decision making. *J. Exp. Soc. Psychol* **1997**, *33*, 345–366. [CrossRef] [PubMed]
19. 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*; Springer: London, UK, 2007; pp. 231–239.
20. Cyra, L.; Gorski, J. Supporting Compliance with Security Standards by Trust Case Templates. In Proceedings of the 2nd International Conference on Dependability of Computer Systems (DepCoS-RELCOMEX 2007), Szklarska Poreba, Poland, 14–16 June 2007; pp. 91–98.
21. Hawkins, R.; Kelly, T.; Knight, J.; Graydon, P. A New Approach to Creating Clear Safety Arguments. In *Advances in Systems Safety*; Springer: Berlin, Germany, 2011; pp. 3–23.
22. Wang, R.; Guiochet, J.; Motet, G.; Schön, W. D-S Theory for Argument Confidence Assessment. In Proceedings of the 4th International Conference on Belief Functions (BELIEF 2016), Prague, Czech Republic, 21–23 September 2016; pp. 190–200.
23. Derntl, M. Basics of Research Paper Writing and Publishing. *J. Technol. Enhan. Learn.* **2014**, *6*, 105–123. Available online: http://dbis.rwth-aachen.de/~{}derntl/papers/misc/paperwriting.pdf (accessed on 20 January 2019). [CrossRef]