SN Computer Science (2021) 2:443
https://doi.org/10.1007/s42979-021-00832-0

ORIGINAL RESEARCH

Expertise Detection in Crowdsourcing Forums Using the Composition of Latent Topics and Joint Syntactic–Semantic Cues

Yonas Demeke Woldemariam[1]

Received: 23 February 2021 / Accepted: 23 August 2021 / Published online: 3 September 2021

Abstract

We develop an NLP method for inferring potential contributors among multitude of users within crowdsourcing forums (CSFs). The method basically provides a way to predict expertise from their structures (syntax–semantic patterns) when crowdsourced votes are unavailable. It primarily deals with tackling core adverse conditions, which hinder the identification of crowds' expertise levels, and standardization of measuring linguistic quality of crowdsourced text. To solve the former, an expertise estimation and linguistic feature annotation algorithm is developed. To approach the later, a comprehensive linguistic characterization of crowdsourced text, along with extensive joint syntax–punctuation analyses, have been carried out. The entire corpora are comprised of approximately 8 different domains, 3 million and 50,000 sentences, and 32 million and 90,000 words, contributed by a crowd of 50,000 users. The analyses revealed six major linguistic patterns, identified on the basis of ordered lists of structural (syntactic) categories, learned from grammatical constructions, practiced by major groups of experts. In addition, nine different text-oriented expertise dimensions are identified, as crucial steps towards establishing standard linguistic-based expertise-framework for most CSFs. Potentially, the resulting framework simplifies the measurement of crowds' proficiency, in those particular forums, where crowds' tasks (e.g., answering questions, technically discerning deep features within images of galaxies for classifying them into certain categories) are intimately connected with their writing (e.g., describing answers illustratively, expressing complex phenomena observed in classified images). Moreover, wide varieties of linguistic annotations: latent topic annotations, named entities, syntactic and punctuation annotations, semantic and character set annotations, word and character n-grams ($n = 2$ and 3) annotations, are extracted. That is for building baseline and enhanced versions of expertise models (about 20 different models built). The successive achievements of enhancing baseline models, with iteratively adding linguistic feature annotations in a two-stage enhancement process, indicate the adaptability of the learned models.

Keywords Expertise modeling · Latent topics · Syntax–semantics

Introduction

Crowdsourcing forums (CSFs) are typically characterized by quite mixed crowds of *expert* and *non-expert users*. In addition, they are also open to any participants interested in sharing expertise in various fields. Thereby, **effectively discerning expertise hidden in crowds' contributions** takes to look into individual aspects of their overall contributions. Especially, when their contributions involve *writing* or is intertwined with *natural text*, whose relationship with the associated *proficiency in core tasks*, needs to be clearly understood. Users' contributions vary with types of forums. For instance, in forums like *community question answering (CQA)* (e.g., StackExchange (SE)[1], Quora[2]), users contribute answers for asked questions. However, in forums like *crowdsourcing research projects (e.g., Planet Hunters TESS*[3], *GalaxyZoo*[4], participants contribute *classification annotations* of various images, along with text describing those images. Besides *text acquisition* from the crowd, some of these forums explicitly reflect their concern for linguistic quality. To achieve that they provide *text writing guidelines*[5].

✉ Yonas Demeke Woldemariam
yonasd@cs.umu.se

[1] Department of Computing Science, Umeå University, Umeå, Sweden

[1] https://stackexchange.com/.
[2] https://www.quora.com.
[3] https://www.zooniverse.org/projects/nora-dot-eisner/planet-hunters-tess/.
[4] https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/.
[5] https://stackoverflow.com/help/how-to-answer.

Despite such quality concern, and the fact that there is a wealth of *crowdsourced textual content* in these forums, it never appears to be exploited for rating **user expertise**.

Identifying *experts* among multitude of users in *text-intensive CSFs* requires to critically evaluate the *quality of crowd's core tasks*. In addition, that needs to be followed by deeply consider *internal linguistic constructions* of crowds' text. However, in most crowdsourcing forums, for instance, in question-answering forums, the expertise or the proficiency level of users is estimated by solely vote counting. But there is a possibility of combining such meta-information with the linguistic quality of answers/question content to compute the overall competence.

That basically involves clearly understanding the underlying connection between the core tasks and the associated text [1, 2]. Followed by a good *linguistic representation and annotations* of the text for *establishing a strong computational link* between the two variables. Nevertheless, as non-linguistic meta features (e.g., number of votes) typically characterize most emerging crowdsourcing sites. That seems to lead many predictive analytics relying on such *simple surface features* to evaluate the quality of users' contribution, while *ignoring linguistic quality* of users' textual content [3]. Such predictions, however, could advance and benefit a lot through the exploitation of textual content and capturing their *complex linguistic constructions* [4, 5].

The Scope and the Focus of the Study

In this study, we explore the possibility of detecting users' expertise from their text and attempt to answer the research question stated below. The study targets users in crowdsourcing forums, as they present potential challenges which have been discussed in detail in the later sections.

> **The primary research question:** To what extent does linguistic information (particulary, joint syntactic-semantic and latent topics) detect and predict user expertise in crowdsourcing forums?

Our study also makes the following major contributions:

1. The development of an expertise estimation algorithm for crowdsourcing users, whose pseudocode has been presented in Algorithm 1.
2. The identification of distinctive linguistic patterns characterizing crowdsourced text, particularly **joint syntactic–punctuation patterns**. That have been achieved through extensive analysis of grammatical constructions of various expert groups.
3. The identification of core syntactic categories that potentially define or estimate expertise in crowdsourcing forums.
4. The development of computational linguistic models that predict users' proficiency from syntactic structures of text. Existing studies on crowdsourced forums focus on surface meta-features or shallow linguistic features to predict various information. However, in this study, we look into internal grammatical structures of text, which helps improve baseline models and achieve accurate predictions.

Motivation, Challenges and Application Areas

Here, we just wanted to briefly touch up on the motivation of our current study and its potential application areas. While we provide the details in the next paragraphs why it is important to apply NLP, the major reason is we are targeting particularly the textual content part of crowdsourcing forums. Therefore, NLP seems to be a plausible choice to look into such content.

Given that there are ever emerging semantic technologies and growing efforts to advance web content searchability. Recently, cross-media analysis platforms such as MICO [6] and EUMSSI [7] have been released. During the early stages (first milestones of the MICO project) of the MICO's platform architecture, we contributed text analysis components as a project partner [8]. In the later stages, we explored the adaptation of language specific components for less-resourced languages [9, 10]. The aim, in former case, is to make crowdsourcing content searchable based on users' sentiment, while the later case aims to widen the accessibility of audio/video content in less-resourced languages.

The method presented in Ref. [8], could be extrapolated to other media platforms, where user reviews management is quite constrained. Considering media reviews, some of them are very short and even less useful. Some others are very important and characterized by technical passages, valuable named entities and URL links. For instance, the comment sorting algorithm (web-interface) of Youtube, severely limited to only a couple of options (basically recency). However, many other sorting and searching options including sentiment and other either content-related attributes (e.g., credibility, informativeness), author-related attributes (e.g., expertise, reputation) are interesting. We realize that such multimodal searches take a huge effort of applying and building dedicated NLP-based utilities. That is for filtering comments and providing prediction services. The sentiment analysis component within the MICO platform [8] has been integrated as a meta-data extractor; and implemented as an analysis chain that includes three sub-components (tasks): chat message pre-processor (aka a chat room cleaner), natural language processing (e.g., part-of-speech tagging, named

entity recognition) and sentiment annotations. In this current study, we aim to explore another potential annotation associated with authors, namely expertise or proficiency. That plays the role of enhancing and widening content searchability. In that case, it needs to go beyond sentiment analysis, looking into internal linguistic constructions of users' text and characterize authors with important expertise dimensions.

Regardless of the types of CSFs, there is quite a common interest of making clear distinction among experts based on the quality of their contributions. Thus, an orchestrating mechanism for systematically harmonizing between *users' expertise* and their associated *text*, needs to be developed. That is to be achieved by taking the quality of core tasks (e.g., programming in SE StackOverflow (SO), classifying images in Zooniverse Galaxy Zoo), and the associated *textual content quality* into account. Putting the mechanism in action, eventually helps for effectively embracing the dualism of the actual core tasks along with the text. That also implicitly encourages users to contribute high-quality content in the forums. Then, ultimately, which turns out to improve the overall quality of the forum, as users' text often comes into the possession of the forums.

The intended mechanism does not exist at all in types of forums like Zooniverse. However, there exist user rating systems within some CSFs such as *the point system in Yahoo!Answers* and *the reputation system in SE*. Such systems operate based on **solely vote counting. Yet, which might signal more of content's popularity than content's quality**. Moreover, the *actual textual content part* seems to be perceived as a trifle and cut off from the equation of contribution assessments. Apparently, that self-contradicts with their stated guidelines (e.g., in SE[6]) for achieving standard content quality. Yet, that failed to have dedicated utilities that directly deal with textual content quality. Consequently, that turns out to affect the credibility of the forums as a result of paying no attention for linguistic aspects. That also causes to leave poor quality content on their platforms. Nevertheless, every score assigned by these systems could be regarded as a seed to be brought forth from their *majority vote scheme logics*, as most social media platforms do, so to speak. However, this score has a potential to germinate into a *near-realistic expertise score* via the light of *concrete linguistic evidence* emanating from the knowledge present in the attached text.

The Selected Crowdsourcing Forums

In this study, world leading and largest crowdsourcing forums (Zooniverse and StackExchange) have been selected. In contrast with other forums, these forums satisfy a couple of conditions. First, user activities co-exist with their writings. Second, since they are quite old they have got a reasonable amount of crowdsourced text for our study. We provide a short background about such forums as follows.

Zooniverse: as a Crowdsourcing Scientific Research Forum

Zooniverse hosts about 50 different projects to carry out *crowdsourced research works*. Thereby, approximately 1.6 million volunteers with various expertise or skills. Volunteer users share their experiences by analyzing, classifying and describing images of various kinds.

Nearly 518 million classifications have been made by the volunteers, along with a large number of discussion posts. Nevertheless, no tool has been implemented yet, that guarantees whether those classifications are right or wrong. Because of that, expertise levels of *user classifiers* remain unknown. However, the available *crowdsourced text* generated from the discussion posts can be systematically merged together with classification information. That could be taken as an important step towards estimating users' proficiency. To achieve that, among active Zooniverse's projects those which are old and having sufficiently enough data are used in this study. These are Galaxy Zoo and Snapshot Serengeti projects.

Among all Zooniverse's projects, Galaxy Zoo is the oldest as well as the biggest. Which is hosted by Oxford University and the Adler Planetarium, to study the magnificence design and fabrics of the observable physical Universe. Galaxy Zoo has over 49,000 users, who made nearly 1.5 million classifications, of over 212,000 subjects (images) captured by the Hubble Space telescope. On the other hand, the citizen science project Snapshot Serengeti explores the dynamics of wildlife living in the Serengeti national park. Within the project, there are over 2.5 million classifications of about half million subjects, contributed by approximately 14,000 registered users.

StackExchange: as a Crowdsourcing Question-Answering Forum

Question-answering (QA) forums (aka community QA) provide knowledge-sharing sites, where different types of questions can be asked and answered in written form. Usually such questions are collaboratively answered by multiple users with the possibility of having varied answers. In addition, the resulting answers are made available for public use (possibly for further improvements) and shared among other on-line communities. CQA content is also indexed by search engines where questions can be asked in the form of queries. In most cases, well-formulated questions receive many good answers. Similarly, well-written (in terms of grammar

[6] https://stackoverflow.com/help/how-to-answer.

and structure) answers tend to be voted up by many users and top-ranked. That, eventually, leads users to gain high reputation (expertise scores) within a particular CQA forum. On the other hand, questions/answers that are vague and off-topic remain unanswered/voted-down and even deleted from CQA sites.

Examples of CQA sites include Stack Exchange (SE), Yahoo!Answers (YA)[7] and Quora. In comparison, SE and Yahoo!Answers are more diverse (in topics) as well as older and larger. On the other hand, Quora is multi-lingual and recent. Regardless of their credibility, they are increasingly becoming important (re)sources of information for various types of open-ended questions. That cover a wide range of topics from simple programming to complex life questions.

The StackExchange network has kept growing since its release. Currently, it contains about 176 different topic-oriented forums along with over 12 million users. Reportedly claimed to have over 1 billion page visits as well. Among them, the StackOverflow forum is the largest (in terms of number of users, as well as asked and answered questions). It also has a unique content feature of having *natural text* and *code snippets* together. Together with StackOverflow, other five forums have been included in this study. Three of them (Server Fault, Super User and Ask Ubuntu) are chosen based their relatedness with StackOverflow. In addition to that, the number of questions and answers along with answer rate (i.e., nearly 100% answer rate) are considered. The remaining two forums (Mathematics and English) pretty much satisfy the same criteria as the three forums, except their unrelatedness with StackOverflow. Such deviation (unrelatedness) is important to evaluate our method on out-domain datasets and make a good conclusion.

Within StackExchange, the reputation system assigns scores for users for their contribution. Likely, high competent users contribute higher quality content than less competent ones, both content wise and language wise. In other words, we assume that content posted by highly proficient users tend to be technically valid (meaningful), theoretically grounded (motivated with theories and good examples) and grammatically sound. Content posted by low-level users, however, are characterized by wrong and less-focused answers, incorrect grammar and punctuations. However, the mechanism (used by the majority of CQA forums) employed to assess proficiency and make such distinction between users is constrained in many ways. Moreover, there seems to exist the tendency that answers provided by top-level users get accepted quickly and frequently. Yet, such mechanism might wrongly marked such users as "reputable" and their answers are factoid [11–13]. Thus, the whole quality problem seems to lie in the vote counting mechanism.

[7] https://answers.yahoo.com/.

Expertise in the Context of CSFs

The notion of *expertise* or *proficiency* is quite broad and multidimensional. It is often perceived in many ways unless restricted with context. It always entails a number of aspects (e.g., core skills, moral ethics, cognitive, authoritative(ness), assertive(ness), and communicative) whose perception and measurement is not straightforward. As a result, the accuracy of *overall expertise assessment* is not only contingent on the quality of each aspect, but a good understanding of them.

It gets even more complicated when comes to *open volunteer-based CSFs*. where *experts* and *non-experts* could register without any known prior skill requirements, and whose participation is *loosely controlled*. That is fundamentally due to a couple of reasons. First, there is no existing *concrete conceptual framework* providing a clear understanding of what *expertise* in such forums needs to entail in general. In addition, the connection with *users' text* in particular. Second, there is no established standard to precisely measure the strength of the relationship between actual tasks (e.g., programming, classification) and the associated text quality.

In this study, we consider two types of **expertise in context**: *proficiency in classification (PIC)* and *proficiency in programming (PIP)*. We provide their definition and core characteristics, followed by their estimation.

Proficiency in Classification (PIC)

Within the science crowdsourcing Zooniverse, *PIC* could be regarded as an overall measure of the quality (of being correct) of *image classifications* made by a particular *volunteer expert*. Users are guided by flow charts containing hierarchical questions. That lead users to classification decisions. Effective classifications, for instance classifications of telescope images of galaxies, could be accomplished, and always involve three major consecutive tasks. These are users' subtleness and active cognition or a close observation for identifying hidden features in images (e.g., planetary transits, light curves), reasoning and analysis, and labeling with certain categories.

Interestingly enough, perhaps, such cognitive processes have a chance of being mirrored in the subsequent writing. That leads to establishing important connection between the quality of the actual *classification task* and the associated *writing*. An illustrative sample snapshot of a galaxy image description is provided in Fig. 1.

Likely, as experts get more experienced, their cognitive processes get more sharper to discern complex phenomena. For example, deeply examining the juxtaposition of clustered stars, might lead to accurately classifying them into right classes of galaxies. Essentially, their discernment ability goes further to give profoundly visual descriptions as vivid as Space telescope

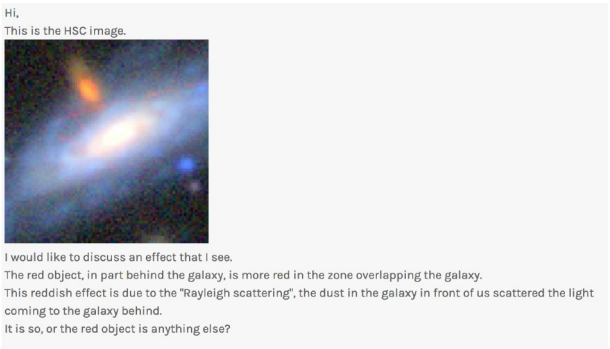


Fig. 1 A piece of **crowdsourced text**, contributed by a Galaxy Zoo expert to **discuss**, possibly after **classifying an image of galaxy into certain categories**

captured images. The resulting descriptions and expressions do not only give pictorial accounts of the inevitable magnificence beauty of the Universe. But they tend to make clear distinction between *expert* and *non-expert users*. Perhaps, average and low level experts tend to use quit common language, evidentially, accompanied by high frequency of interjections. In addition, their expressions are pretty much terminated by exclamation marks as a fulfillment of their emotion. To the contrary, high-level experts' words tend to be more imagery, graphical and powerful enough to express the reality of (behind) the observed images.

Unlike other forums (e.g., StackExchange) which apply well established *point systems*, Zooniverse lacks *quality measures* for assigning *PIC scores* and the *ground truth* for determining expertise levels of users. Thus, we implemented an algorithm (shown in Algorithm 1) that computes *PIC scores* and build gold-standard data based on the *weighted majority votes scheme*.

In this study, we attempted to answer, what defines a **good crowdsourced text expression**, through an extensive **syntax–punctuation pattern analysis**. As a result, 9 **expertise dimensions** related with linguistic qualities (e.g., completeness, descriptive, and analytical) have been identified. For the good characterization of the content, we considered *syntactic structures together with punctuation*. That is because they allow to capture **larger semantic units** which could be derived from **larger syntactic categories (e.g., noun and verb phrases)** compared to individual words whose meanings always leads to perplexity. Yet, word-based methods (e.g., word *n*-grams, topic modeling, and character *n*-grams) have been used to build our predictive models.

Algorithm 1: Proficiency in Classification (PIC) Scores Computation

Input: Subject (Image) Profiles (subPro) and User Profiles (userPro)
Output: userProficiencyMap<userId, proficiency>
initialization
    `// build a map from subPro with <subjectId, classCountList> pairs`
$HashMap < String, List < Int >> subjectMap \leftarrow null$
    `// omit var. intialization for the sake of space`
while *(null != (line = subjectReader.readLine()))* **do**
    $tokener \leftarrow line$
    $subject \leftarrow tokener$
    $subjectId \leftarrow subject.getString(\text{“}subjectId\text{”})$
    $totalCount \leftarrow subject.getInt(\text{“}sum\text{”})$
    $classACount \leftarrow subject.getInt(\text{“}classA\text{”})$
    $classBCount \leftarrow subject.getInt(\text{“}classB\text{”})$
    $classCCount \leftarrow subject.getInt(\text{“}classC\text{”})$
    classCountList.put (totalCount, classACount, classBCount, classCCount)
    subjectMap.put(subjectId,classCountList)
`// compute proficiency scores using subjectMap`
$HashMap < String, Int > userProficiencyMap \leftarrow null$
$userReader \leftarrow userPro$
`// a series of questions for each subject being classified`
$questionSet[] \leftarrow Q_1 \ldots Q_n$
while *(null != (line = userReader.readLine()))* **do**
    $tokener \leftarrow line$
    $classification \leftarrow tokener$
    $subjectId \leftarrow classification.getString(\text{“}subjectId\text{”})$
    $classCountList2 \leftarrow subjectMap.get(subjectId)$
    `// check that the subject is classified by at least 10 users`
    if $classCountList2 != null$ **then**
        if $classCountList2.sum \geq 10$ $annotations \leftarrow classification.getAnswerArray(\text{“}annotations\text{”})$
        $answer \leftarrow null$; $j \leftarrow 0$; $found \leftarrow false$
        for $i = 0$ *to annotations.length()* **do**
            $question \leftarrow annotations.getQuestion(i)$
            while $j < questionSet.length$ ***and*** $!found$ **do**
                if $question.has(questionSet[j])$ **then**
                    $answer \leftarrow question.getAnswer(questionSet[j])$
                    $found \leftarrow true$
                $j \leftarrow j+1$
        if $answer != null$ **then**
            $classifiedSub \leftarrow classifiedSub + 1$ `// assign a proficiencyScore based on weighted majority votes`
            if $answer.equals(\text{“}classA\text{”})$ **then**
                $proficiency \leftarrow proficiency + classCount.classA/classCount.sum$
            else if $answer.equals(\text{“}classB\text{”})$ **then**
                $proficiency \leftarrow proficiency + classCount.classB/classCount.sum$
            else if $answer.equals(\text{“}classB\text{”})$ **then**
                $proficiency \leftarrow proficiency + classCount.classC/classCount.sum$
$proficiency \leftarrow proficiency/classifiedSub$ **if** $classifiedSub \geq 10$ **then**
    userProficiencyMap.put(userId, proficiency)
return userProficiencyMap

Proficiency in Programming (PIP)

PIP has pretty much a standard definition, in contrast to *PIC*. That simply refers to a measure of competence in writing computer programs. Nevertheless, as we are considering programmers or any type of users in StackOverflow (SO), the definition needs to be restricted to fit what it means in there. Within SO, the closest metric for measuring users' performance is **reputation** [11, 14, 15]. A cumulative reputation score received by a particular user could be perceived as a measure of an overall quality of the contribution for answering and asking questions. It also involves editing questions and answers. Either the questions or the answers do not necessarily be associated with *coding*. Therefore, we take *the notion of PIP* in its loose sense. A description on an estimation of *reputation scores* is provided in Ref. [5].

Reputation systems treat various groups of users differently. For example, *answerers*, particularly those whose answers get *accepted* possess much reputation scores than *askers*. Moreover, due to the fact that *answering* alone seems to be self-evident that users are demonstrating their programming skills. Therefore, our focus shifts towards those *answerers*. Likewise, many votes for an answer implies inter-agreement between other users for its certain quality. Eventually, the acceptance of that particular answer approves its being best among other alternative answers. To successfully pass such two important major steps, any answer needs to meet certain programming qualities (e.g., readability and having illustrative code snippets) and linguistic qualities (e.g., completeness, descriptiveness). However, there might be subjectivity among users rating the answer. The subjectivity caused by the vote-based rating system, obviously, raises many potential issues on the validity of the users' competence evaluation method.

Possibly, the subjectivity can be reduced via the systematic exploitation of the two parts of such programming and natural text carrying content. The former can be captured with *character sets information* present in the code snippets and the non-linguistic meta-data features (e.g., vote counts). The later can be captured through *natural text analysis methods (e.g., syntactic and semantic parsing, topic modeling)*. Having such rich linguistic and non-linguistic annotations, potentially allows the *detection of expert levels of users*. That also improves the quality of the estimation of users' proficiency.

Related Literature Review

There are several interesting works on text analysis or social media analytics [4, 15–19]. However, studies which directly target CSFs for expertise detection do not seem to exist very much. Some provide insights into dynamics of user-generated reviews and ratings that potentially lead to the development of integrated frameworks, for instance, Yung-Chun et al. in Ref. [19] proposes a pipeline of a data crawler, data pre-processing and visual analytics that improves the quality of sentiment analysis. Others reveal authors related information from various sources of textual content. One way or another, most of them are important for the development of expert detection methods.

In this section, we cover related literature from three major areas: **information quality**, **correlational studies between text and related tasks** and **psycholinguistic analysis**. Following brief summaries of the core part of the methods and findings discussed in these areas, we describe the context of our study at the end of this section.

Methods on the evaluation of content quality within CSFs helps set *ground truth* for characterizing and measuring users' expertise. That is on the basis of certain established metrics of content quality. Quality appears to be quite problematic in crowdsourcing forums. That is mainly due to the fact that they are highly marked by *loose control of user-generated content*. To address this concern, various assessment methods have been developed. For instance, Zhu et al. in Ref. [16] proposes a model to evaluate the quality of user-generated answers in CQAs. The model has been presented as a multi-dimensional framework containing 13 answer-intensive criteria (e.g., informativeness, relevance). Such criteria are collected from CQAs user opinions as a form of a questionnaire and to be judged by human raters. Their method of building a ground truth on the basis of such subjective opinions is to tackle the subjectivity of answers' quality. That, however, raises some concerns, while it is an important step towards defining answer quality in itself. Among these criteria, only one deals with the expertise of users (answerers) to guarantee whether the answer is provided by experts. In addition, the majority of them have been set from answers' quality point of view. Nevertheless, we noticed that the resulting framework presented appears to be erroneousness in estimating the overall quality measure. Perhaps, that occurred during assigning weights to the identified quality dimensions.

The built model in Ref. [16] has been applied by other content-quality studies within CQAs. For example, Shah and Pomerantz in Ref. [15] adopt the model to predict the quality of answers by combining the 13 criteria defined in Ref. [16]. But they added two more sets of criteria. Thereby, we are particularly interested in how they detect whether answers are provided by experts. Their approach also gives other valuable insights on possible ways of classifying user-generated answers within in CQAs. In both studies, the presence of *technical (professional) terms* within answers appear to be an indicator whether the answers are provided by experts. As both involve human annotators for rating answers (question–answers pairs) with certain scales,

they tend to be subject to *human bias*. Despite the bias, the authors in Ref. [15] argue that their *inter-rater agreement assessment* proves that there is quite common consensus among the raters on the evaluation of expertise levels of answerers. Probably, the *binarization* of their scale of 1–5, prior to the inter-rater agreement assessment, has helped the reduction of the bias. Furthermore, their logistic regression-based classifier strengthens the claim, yielding about 80% accuracy, though low R-squared has been scored.

The relationship between users' performance and their associated writing has been studied from psycholinguistic perspectives. How language use within a group of cockpit staff (pilots and crew members) is possibly connected with their performance has been explored in Ref. [2]. They use the text analysis method developed for deriving psychological meanings of words by Tausczik and Pennebaker. The authors analyzed speech transcriptions extracted from a cockpit voice recorder. In addition, they found out a significant relationship based on the computed correlation between the two variables. Thereby, how staff' inquisitiveness (asking habit) varies across their positions. That seems to decrease as position increases. According to their study, authoritativeness appears to mark the language used by high-performance staff. On the other hand, self-focused expressions (evidenced by a high percentage use of first-person singular pronouns) seem to characterize the language patterns of low-position staff members. Authoritativeness in Ref. [1] perceived as a high use of first-person plural and a less use of first-person singular pronouns. Nevertheless, in reality, that is not always true. That has been used as an important user related variable, together with other three variables (e.g., reputation and social connectedness), to define the quality of questions and answers within the MathOverflow forum [18]. Nevertheless, it happens to be the least significant predictor of both question and answer quality, while reputation appears to be the most dominant factor.

Methods for inferring experts' performance have been also studied as a function of simple to complex linguistic features. That are extracted from wide varieties of textual data. Medical reports, for instance, have been used to assess the competence of medical students [4]. The authors identified core competence dimensions via the support of domain-oriented resources (a unified medical language system and a knowledgeMap concept indexer). The authors transformed the students' medical reports into bag-of-word representations. That followed by training SVM, logistic regression and naive Bayes classifier. A corpus of wine reviews written by wine experts, has been also studied by Croijmans, et al. However, the authors aim to predict various properties (e.g., grape variety, color) of wine, instead of authors' expertise. The authors train SVM classifiers on linguistic information extracted from about 76,410 reviews. Largely, the information contains part-of-speech tags and meta-data about wine types, and reportedly found 95.5% F-score and 3% accuracy.

Successive works [5, 8, 20] have been done to predict expressed sentiments. In addition, these works also aim to discover rules governing the conjectured link between users' performance and their associated text posts. The authors attempted to progressively understand the underlying link through numerous tasks. That include performance estimation, linguistic feature extraction and statistical model construction. The results of their early study Wolemariam et al. 2017 provide foundational insights, precisely on the computational relationships between the selected linguistic information (e.g., syntactic, bag-of-words, and punctuation) and competence scores. In addition, the authors also provided empirical evidence for answering some basic questions concerning relevant text representations and linguistic features. Essentially, they also argue why their methods have been constrained. Moreover, they identified potential areas of CSFs where their methods could be so effective and practical. Thereby, for instance, the syntactic representation appears to yield best predictions on the trained classifiers, compared to the bag-of-words representation. In addition, their competence estimation method as well as the actual textual content have been found to be limited in many ways.

That eventually led them to another related study [5], in which the authors extend and improve their previous approaches. That were in terms of methods for users' competence annotation, text representations and modeling techniques. In Ref. [5], Woldemarim argues that existing methods in CQAs for evaluating users' reputation are incomplete as they fail to incorporate linguistic quality present in actual user-generated textual content. How such potential failure, ultimately affects the quality of the entire CQA forums described in detail in Ref. [5]. The validity of their approach as well as the usefulness of their resulting models have been evaluated on with fairly large and varied datasets. As a result the authors conclude that users' performance is subject to variations of linguistic information, to the degree of 80% of R-squared with 3% errors.

In our current study, we aim develop an NLP method that primarily involves identifying typical **structural linguistic patterns** along with **core expertise dimensions**, and designing an **expertise-oriented conceptual framework**. That potentially, helps somehow *generalize* the *linguistic characterization* of **crowdsourced text** with respect to the corresponding measured **expertise**. That would *simplify* the computation of expertise within CSFs, as it allows to explore how structural and topical information variations influence crowds' expertise levels. To achieve the generalization, world's largest crowdsourcing forums have been targeted. We provide a comprehensive **joint syntax–punctuation**

characterization via **syntactic parse tree analysis**, to map into domain-oriented fine-grained **expertise dimensions**.

Moreover, wide varieties of linguistic annotations: *latent topic annotations*, *named entities*, *syntactic and punctuation annotations*, *semantic and character set annotations*, *word and character n-grams (n = 2 and 3) annotations*, are extracted. That is for building *baseline* and *enhanced* versions of *expertise models*. As the resulting expertise models are intended to be validated and evaluated on quite miscellaneous (in-domain, related domain and out-domain) sources of crowdsourced text, possibly *domain generalizability* and *adaptability* would be achieved.

Methodology

In this section, we discuss the major NLP tasks carried out for accomplishing the entire research design and experimental setting. That basically involves descriptions of methods for preparing corpora used for model building and evaluations. In addition, methods for document representation and characterization with the selected linguistic annotations are discussed along with models training, validation and evaluation.

Experimental Data and Pre-processing NLP Tasks

The experimental data contain representative textual data written by nearly 428,200 different authors. That is collected from wide varieties of topics in the area of CQA and citizen science projects.

The Zooniverse's corpus: contains a unified parallel data-dumps (datasets) of *Galaxy Zoo* and *Snapshot Serengeti* citizen science projects. The corpus is collected from the internal servers of Oxford University, where such projects are hosted. The size of each dataset is quite insufficient and small to represent the crowdsourcing domain compared to SE. Unifying the two datasets, thus, has been regarded as a reasonable domain adaptation strategy [21, 22]. Such strategy is applied after preliminary experiments on each dataset. Nevertheless, it would be worth trying other neural-based adaptation methods [23, 24]. That is particularly developed to solve resource scarcity problems in speech recognition. The resulting textual content in the corpus has been contributed by over 13,000 distinct volunteer users who made about 10 million classifications. The entire textual data are comprised of around 50,000 sentences with the mean sentence length of 5 and 90,000 words. Following randomly (with shuffled sampling) partitioning the data into three subsets: 70% for training, 10% for validation (development), and the remaining 20% for evaluation sets.

The StackExchange's corpus contains six different datasets from various forums within the SE network:*StackOverflow, ServerFault, SuperUsers, Ask Ubuntu, Mathematics and English*. These datasets have been directly collected from SE[8]. Unlike Zooniverse, where a prior agreement might be needed as the data are not open to the public yet. SE regularly loads data archives on its repository. The corpus has been made to contain only answer-content to reflect users' expertise better, instead of including all types of posts (e.g., questions, comments). It is also attempt to further render expertise knowledge and ensure the best quality of the data. Therefore, only users ranging with reasonable expertise levels (users answering at least five questions) to prodigies (e.g., exceptional users answering more than 1000 questions along with up to 80% of their answers have been first ranked) are included. SE comprised of three in-domain and related domains, and three independent out-domain corpora. Approximately, over a half million answers from a total of 27,000 users are included. The entire text corpus contains 3 million sentences and 32 million words (around 8% of them are unique words). For SE, the same division pattern has been followed as the Zooniverse corpus.

Extraction of the below linguistic features is performed after the following pre-processing and text analysis tasks. The extraction aims to generate plain texts using text-processing Java libraries and text analysis packages.

Data acquisition and cleaning: involves turning repository XML data-dumps into database tables via MongoDb database and python scripts. Then, followed by removing stop words (only for document vectors generation, have been kept for parse tree generation), noisy features such as XML tags, and filtering terms ranging a word length of (2–4).

Tokenization: splits crowdsourced text into a sequence of words (tokens) via non-alphanumeric characters.

Stemming: involves generating the morphological root forms (morphemes) of the tokenized words. That is done by stripping their suffixes or prefixes off, using the Porter stemming algorithm [25].

Part of speech tagging: involves labeling individual words with part-of-speech tags (e.g., verbs, nouns)

Named entity recognition: detects names of people, places and so on, mentioned within text.

Syntactic (constituency) parsing: parses input text and generates constituency (phrase-structured) parse trees,

[8] https://archive.org/details/stackexchange.

as shown in Fig. 6a, from which structural information can be extracted.

Semantic (dependency) parsing: builds dependency graphs from input text and extracts dependency relations, as shown in Fig. 6b.

Latent topics extraction: performs the task of topic annotations of input text via latent topic modeling.

Computing TF-IDF scores: includes calculating and weighting word frequencies and their IDFs (inverse document frequency) for each user's document.

Generating a document-term matrix: performs the task of constructing a TF-IDF matrix using the calculated TF-IDF scores (in the above step).

Extraction of number of tokens: returns the frequency counts of tokens in each document and is a very important feature in probabilistic models.

Extraction of aggregate tokens length: calculates the size of each document by aggregating the frequency counts of all tokens occurring in that document.

Document Representations and Linguistic Annotations

Structural Representation and Punctuation Annotations

We aim to build representative models that could possibly be shared by wide varieties of CSFs. That is regardless of their specific domains. Therefore, a document representation that best fits such goal has been used. That is a *syntactic representation*, where each sentence of a document can be hierarchically represented bottom–up from a word-level, a phrase level, a clause level to a sentence level. By doing that, we avoid (reduce) a *domain dependency problem*. Such problem is pretty much exhibited by most computational linguistics models, precisely those built on the *bag-of-word (BoW) representation*. To achieve this, *constituency (phrase-structured) parsing* [26, 27] has been applied on 8 different corpora explained above. Subsequently, *constituency tags (syntactic categories)* have been extracted from the resulting parse trees.

Once parse trees have been constructed, the following syntactic categories are extracted from their non-leaf nodes constituting the hierarchical structure of the input corpora:

- Lexical categories consists of part-of-speech tags (POS) (e.g., VB (verb), NN (noun), JJ (adjective)). These tags correspond to the leaf nodes in the parse trees that represent the words in the parsed text.
- Functional categories consists of elements that connect syntactic units together (e.g., MD (modal), DT (determiner), IN (preposition)).
- Phrasal categories include various types of phrases (e.g., PP (prepositional phrases), NP (noun phrases), VP (verb phrases), ADJP (adjective phrases)). Each phrasal category has a set of labeled words with POS tags within a sentence.

Syntactic representations of documents help to deal with the very *structure of linguistic constructions* of users' text. That is without being dependent on specific domains' knowledge. Essentially, such representations provide interesting cues how users with similar *expertise levels* structure their messages (as shown in Tables 1, 2). In addition, syntactic-based models are robust due to the fact that varying words having similar functions do not affect the underlying structural representations of users' text. For example, at a sentence level, a document could be constructed with many alternative terms while keeping the main idea of the sentence. While each variation results a different BoW representation, the structure always remains the same as long as the alternative words play the same syntactic roles.

From the named entities recognition (NER) point of view (in a partial sense), syntax also helps to make interesting observations. That is how names (of e.g., people, places) mentioning habits is related with expertise via the syntactic categories **NNP and NNPS** namely, **singular proper noun** and **plural proper noun**, respectively. As shown in Table 2, such categories make a significant distinction between experts in SE and Zooniverse.

Besides syntactic categories, we consider how users punctuate their text to enrich and reinforce the learning of our models. There are many essential overlapping between the extracted syntactic and punctuation information. Therefore, observing consistent pattens of similar or related syntax–punctuation annotations leads better understanding, and affirms their validity. For instance, considering quite similar occurrences and influences (over expertise) of the syntactic category*SBARQ (questions introduced by a wh-word or a wh-phrase)* and the punctuation mark *?*, somewhat helps establish strong linguistic evidence for inferring expertise from linguistic annotations. Moreover, the below join syntax–punctuation annotation pattern analysis reveals a number of occurrences of such types of overlapping.

Analysis of Joint Syntax–Punctuation Patterns Observed from Grammatical Constructions Within the Data

In connection with syntactic annotations, in this subsection, we provide the characterization of the collected crowdsourced text with joint syntax–punctuation features. Considering linguistic constructions from wide varieties of crowdsourced forums helps to well render and establish quite typical patterns present in such forums. In addition, we also *structurally define crowdsourced text*. We discuss major linguistic patterns formed by various possible combinations of syntactic units. That range from smaller constructs

Table 1 Sorted joint syntax punctuation features into linguistic patterns for the 3 expert groups in StackOverflow: low, middle and top

Low	Middle	Top	Level	Description	Expertise Dim.
NP	NP	NP	Phrase	Noun phrase	Completeness
VP	VP	VP		Verb phrase	
S	S	S	Clause	Declarative Cl.	Simplicity
Per.	Per.	Per.	Pun.	Punctuation	Completeness
PP	PP	PP	Phrase	Prepositional P.	Coherency
SBAR	SBAR	SBAR	Clause	Subordinate C.	Completeness
PRN	ADVP	ADVP	Phrase	Adverb phrase	Descriptive-ness
ADJP	ADJP	ADJP		Adjective Ph.	
ADVP	PRN	PRN		Parenthetical	Clarity
FRAG	FRAG	FRAG		Fragment	In-complete-ness
WHNP	WHNP	WHNP		Wh-noun Ph.	Inquisitiveness
Ques.	Ques.	Ques.	Pun.	Punctuation	
QP	HashT.	HashT.		Punctuation	Social Int.
HashT.	QP	QP	Phrase	Quantifier P.	Analytical
X	PRT	PRT	Word	Particle	Clarity
PRT	X	X	Phrase	Unknown	Complexity
Excla.	SQ	SQ	Clause	Inverted Ques.	Inquisitiveness
SQ	Excla.	Excla.	Pun.	Punctuation	Emotion Exp.
AtSym.	AtSym.	AtSym.		Punctuation	Social Int.
SINV	SINV	SINV	Clause	Inverted Sent.	Focus
INTJ	LST	INTJ	Phrase	Interjection	Emotion Exp.
LST	INTJ	LST		List marker	Analytical
SBARQ	UCP	SBARQ	Clause	Direct question	Inquisitiveness
UCP	SBARQ	UCP	Phrase	UCP	Coherency
CONJP	CONJP	CONJP		Conjunction Ph.	
WHPP	WHPP	WHPP		Wh-pre. Phrase	Inquisitiveness
NX	NX	NX		NP-head marker	Complexity
WHA.	WHA.	WHA.		Wh-adj. Phrase	Inquisitiveness
RRC	RRC	RRC		RRC	In-complete-ness
NAC	NAC	NAC		Scope marker	Clarity

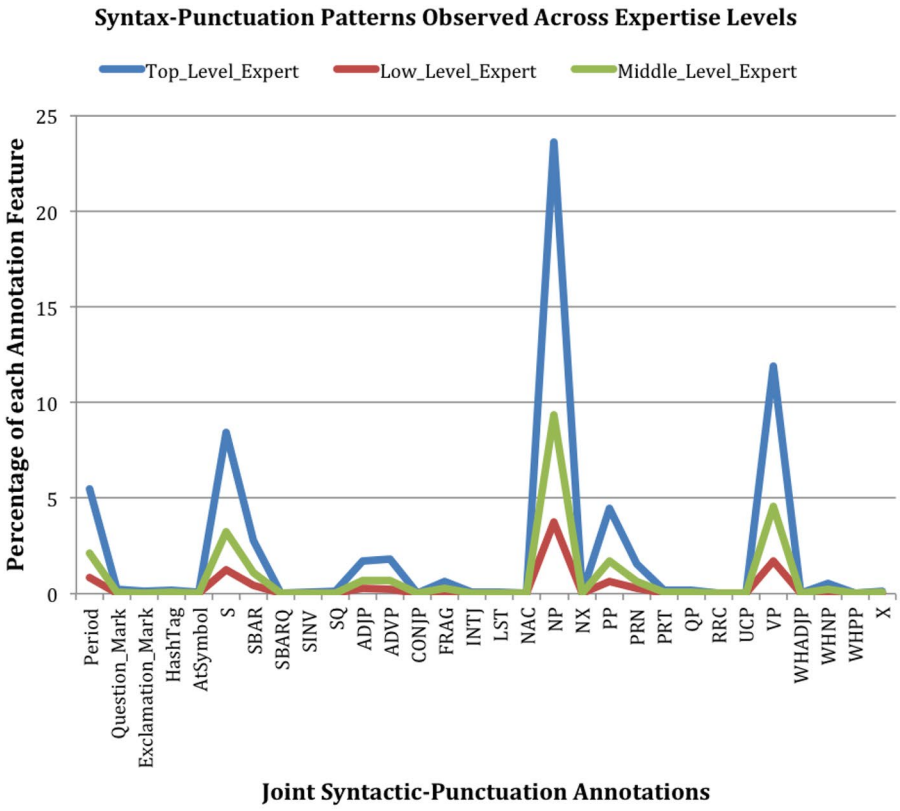
Table 2 Sorted joint syntax punctuation features into linguistic patterns for the 3 expert groups of Zooniverse: low, middle and top

Low	Middle	Top	Level	Description	Expertise Dim.
NP	NP	NP	Phrase	Noun phrase	Completeness
Excla.	Excla.	Excla.	Pun.	Punctuation	Emotion Exp.
VP	VP	VP	Phrase	Verb phrase	Completeness
NN	NN	NN	Word	Noun	
Per.	Per.	Per.	Pun.	Punctuation	
S	HashT.	S	Clause	Declarative C.	Simplicity
DT	S	DT	Pun.	Punctuation	Social Int.
Ques.	DT	Ques.	Word	Determiner	Completeness
JJ	JJ	JJ	Word	Adjective	Descriptiveness
IN	Ques.	IN	Phrase	Punctuation	Inquisitiveness
HashT.	IN	HashT.	Word	Preposition	Coherency
PP	PP	PP	Phrase	Prepositional P.	
SBAR	SBAR	SBAR	Clause	Subordinate C.	Descriptiveness
VBZ	RB	VBZ	Word	Verb	Completeness
NNP	NNP	NNP		Proper noun	Ascertainment
RB	VBZ	RB		Adverb	Social Int.
PRP	PRP	PRP		Personal Pro.	
NNS	VB	NNS		Noun	Completeness
ADJP	ADJP	ADJP	Phrase	Adjective P.	Descriptiveness
VB	NNS	VB	Word	Verb	Completeness
VBP	ADVP	VBP		Verb	Descriptiveness
ADVP	VBP	ADVP	Phrase	Adverb P.	
FRAG	FRAG	FRAG		Fragment	In-completeness
CC	CC	CC	Word	Conjunction	Coherency
WHNP	WHNP	WHNP	Phrase	Wh-noun P.	Inquisitiveness
VBG	VBG	VBG	Word	Verb	Clarity
CD	TO	CD		Cardinal Num.	Analytical
TO	VBD	TO		Infinitive	Completeness
WP	WP	WP		Wh-pronoun	Inquisitiveness
VBD	CD	VBD		Verb	Completeness

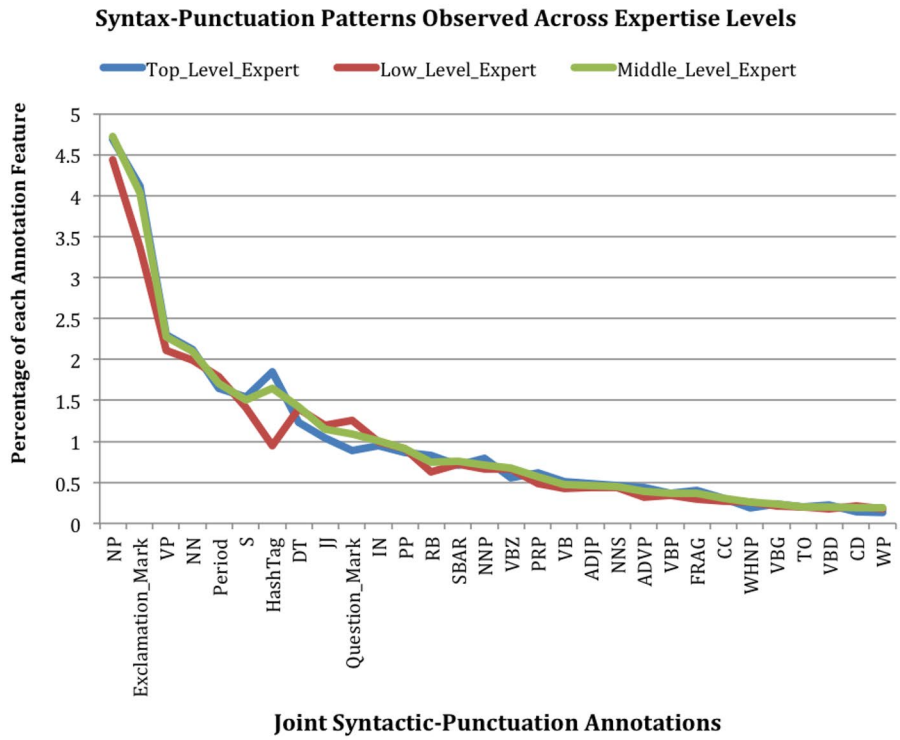
such articles and particles to larger clause and phrase level units such as noun and verb phrases, and simple declarative clauses. Such patterns (e.g., structures of sentences/clauses/phrases/) have been observed from the linguistic constructions practiced by various levels of experts. From each domain, we set aside two groups of users that possess highest and least levels of expertise. Each group covers top or bottom 10% of the entire users from each domain. In case of Zooniverse a group of 1307 users, whereas in SE StackOverflow (2174 users). To see how patterns steadily change over the expertise spectrum as well as reduce the bias probably caused by the marginalization, our analysis embraces the middle-level experts as well.

In the language usage pattern analysis, we achieve three things: identifying common language patterns in CSFs, examining how such pattern changes across expert levels and making distinction between experts. These experts are high-, middle- and low-proficiency scoring experts. We also examine major differences between the two domains in terms of language usage patterns.

As part of the pattern analysis, comprehensive statistic has been generated and summarized in Tables 1 and 2. In addition, the resulting *joint syntax–punctuation pattern*



(a) StackOverflow



(b) Zooniverse

Fig. 2 Percentages of syntactic categories and punctuation marks forming *syntactic–punctuation patterns (SPP)* for 3 groups (high, middle and low) of experts. While in **a NP** and **VP** spike consistently across expert groups, in **b** such categories constituent the top part of the **SPPs' curves**

curves haven been shown in Fig. 2. The two tables clearly show observable key differences and similarities in language usage patterns between the two domains. In addition, each table makes important distinctions between three expert groups (low, middle and high) with in each forum. For instance, looking at the columns *Low, Middle, Top* in each table, signals quite similar rate of using the phrase level categories *NP* and *VP* to form the *NPVP* structure. However, there is an exception use of the exclamation mark which interleaved between *NP* and *VP* in the pattern of Zooniverse. The first two most co-occurring tags form the *NPVP structure*. Nevertheless, the exception in Zooniverse is, perhaps, caused by the nature of the forum, whereby, users' expressions are highly marked by exclamations. Due to some spectacular features observed in the images (e.g., of magnificent galaxies) being discussed by users might drive their emotion and complete their expression by adding exclamation marks. Conversely, expertise-wise, high-level experts in SO are more *analytical* than any other groups of experts. Completely an opposite pattern is noted in *constructing ascertainable expressions* which is evidenced from the usage of (partial) *named entities*, denoted by *NNP* and *NNPS*.

Among other syntactic phenomena closely investigated in this study are the rate at which a sentence fragment occurs. That is denoted as *FRAG* in syntactic parse trees. The investigation also includes the habit of constructing *verb–subject inversion patterns*. While the former makes a clear distinction between those users writing complete sentences or *ill-constructed sentences*, the later signals whether users make any effort for *focus construction* in their writing. Our pattern analysis show that, while experts in Zooniverse make less *ill-constructed sentences* than SO' experts, experts in SO users, particularly high-level experts use much *verb-subject inversion*. However, that does completely capture the entire essence of *focus construction*. Therefore, we apply *topic modeling* which is discussed in the next sub-section. The illustration of such phenomenon has been provided in Table 1, where *FRAG* appears at the upper region of the joint syntax–punctuation patterns. But in Table 2, it appears at a bit lower region. That means, high PIP scoring users tend to use fragmented sentences (with the mean sentence length of 10.48). That is more often than least (whose mean sentence length is a bit longer than the highest scoring experts while shorter than the middle ones) and middle scoring users.

In SO, we analyzed the distribution of the entire *joint syntax–punctuation* information across expertise groups. The analysis is in terms of the number of occurrences of syntactic categories and punctuation marks (i.e., approximately 29.57 million total count). On average, the highest percentage is contributed by high qualified group of users, nearly 64.62% of the entire StackOverflow' corpus. In contrast, the middle and the low-level experts cover the remaining 25.35% and 10.03%, respectively. That generally gives an interesting correlation between levels of expertise and their associated syntax–punctuation usage. Further looking into the distribution of individual feature, precisely the syntactic categories, we observe how the parse trees constructed from each group vary in terms of *types*, *depth* and *complexity*. Almost all *types* of the syntactic categories appear most in the highest group of experts whose trees, in contrast, appear to be a bit *deeper* and *complex* than the other two groups. Zooniverse users, in comparison, contributed less amount of *joint syntax–punctuation* information, i.e., 1.5 million total count. The highest and the middle groups almost equally contributed the most than the low-level group of experts, though the difference is insignificant. The mean sentence length of the groups has reversely ordered with their expertise. In addition, the least, the middle and the highest competent groups is 6.37, 4.19 and 4.75, respectively.

Given the parsed trees along with a punctuation annotation, we identified 30 significantly repeated features composed of syntactic categories and punctuation marks. In addition, they are transformed into a complete ordered list of them. The ordered lists eventually form two distinct linguistic patterns for each forum. To explore how the resulted general linguistic patterns change across the expert groups, we measure the difference between them using the squared Euclidean distance's equation [28] defined below. The distance measurements computed between any two expertise groups show observable differences between major expert groups. The computation is in terms of frequencies of tags constituting linguistic patterns. For instance, the below equation can be applied to calculate the difference between the top and the low experts groups represented by the patterns, $(\text{synTag}_1^T, \ldots, \text{synTag}_n^T)$ or $(\overrightarrow{\text{SynPatrn}_T})$ and $(\text{synTag}_1^L, \ldots, \text{synTag}_n^L)$ or $(\overrightarrow{\text{SynPatrn}_L})$, respectively, where $n = 30$.

$$D(\overrightarrow{\text{SynPatrn}_L}, \overrightarrow{\text{SynPatrn}_T}) = \sqrt{\sum_{i=1}^{n} (\text{synTag}_i{}^1 - \text{synTag}_i{}^2)^2}. \quad (1)$$

The syntactic tags on the list have been also clustered into major *expertise dimensions* based on some existing studies [16, 18]. The studies are on evaluations of user-generated content and contribution quality in CQAs and psycholinguistic studies [1]. Nevertheless, some other dimensions are based on simple pragmatic fine-grained categorization of syntax–punctuation features. Or mapping various related parts of the patterns into possible corresponding expertise dimensions has been done on the basis of perceived intension of authors constructing a particular (sub) pattern. Moreover, decisions were made during the categorization by looking at the similarity of the tags in the function of constructing similar syntactic and semantic units. For instance, all related tags forming noun phrase subtrees get grouped together to define the *completeness* aspect of expertise. Defining the

expertise dimensions is an important step towards establishing a *general expertise-framework* for crowdsourcing sites.

The relationship between the *identified expertise dimensions* and the *joint syntax–punctuation features* have been briefly discussed below. There are some overlapping between such dimensions due to the fact that some tags could play double roles. As some of them just might have only linguistic meanings (than expertise implication) and hard to easily link to specific aspects of expertise. For example, our pattern analysis shows that high-level experts frequently use less *uncommon (unknown) syntactic categories, marked by X* in their sentence constructions than low-level experts. That raises interesting questions and concerns, not easy to exactly figure out what that particular pattern to do with expertise.

Completeness: deals whether users construct complete sentences containing at least *noun phrases (NP)* and *verb phrases (VP)*, along with periods. In addition, other larger and smaller syntactic units (e.g., *subordinate clauses (SBAR)*, *nouns (NN)* and *determiners* (DT)) and tags which could possibly contribute and support the completion are also added into the *composite completeness category*. Conversely, *in-completeness* has been taken as an opposite measure or the *inverse of completeness*, by considering whether *fragments (FRAG)* are present in the parsed text. In connection to that, the use of *reduced relative clauses (RRC)* has also been assumed to partially contribute the formation of in-complete expressions.

Descriptiveness, clarity and simplicity: contribute towards the understandability of the meaning of text. *Adjectives phrases (ADJP)*, *adverb phrases (ADVP)* and *simple declarative clauses (S)* have been considered to form such expertise category. In addition, some other miscellaneous tags (e.g., particles, determiners) have been also added into it, as they play a role of enriching and increasing the size of the content, ultimately making the content to convey a much clear message. For instance, particles alone are not sufficient to form neither a semantic unit nor an expertise category, but combined with other classes of words (e.g., verbs), they form *phrasal verbs* or *larger idiomatic expressions*, and might fit into the *clarity dimension* of expertise. Nevertheless, going further beyond from simple perception, perhaps, would be interesting to exactly get why users prefer those types of expressions.

Analytical, comparison and ascertainment: denote users' ability to provide sound analysis, comparisons and detailed and well referenced information. That are to be evidenced from phrases containing syntactic tags such as *quantifiers (QT)*, *list markers (LST)*, *cardinal numbers (CD)*, *proper nouns (NNP)*, *comparative and superlative adjectives, JJR and JJS, respectively*. The first two tags are present at higher rates within the content of the SO users. In addition, the last two tags are quite common in Zooniverse, precisely appearing consecutively in the *high-level users*. Unlike SO, users in Zooniverse mention peoples names which are evidenced from their usage of *NNPs*, possibly to ascertain various sources of information.

Connecting ideas: roughly reflects how well authors' ideas are joined together. That can be evidenced from the combination of *coordinating conjunctions (CC) and conjunction phrases (CONJP)*, and *prepositions (IN) and prepositional phrases (PP)* and *determiners (DT)*. Within SO, both *CONJP* and *PP* occur at higher rates in high-level experts than middle and low-level users. In contrast, while *CC* has been more frequently used in the high-level experts than other groups of users, *DT* occurs most in middle-level experts within Zooniverse.

Inquisitiveness: shows the asking habits or curiosity of users as a positive indicator of expertise quality. Such dimension has been organized using question marks, syntactic tags (e.g., SQ, WHNP) showing explicit as well as implicit questions in the parsed text. Slight distinguished differences can be observed in the use of these features across the expert groups of SO. For instance, the top expert group asks more than the low one, while these features appear in similar patterns' positions, quite consistently for all groups, though differ in their weights.

Emotion expression: denotes a practice of imparting observational or experiential knowledge. That shows the quality of authors' text conveying emotions. That is, regarded as an essential part of the overall expression of the authors' text, particularly very interesting in Zooniverse due to the *magnificence nature* of the images showing up at displaying scenes, and being discussed. Any polarized (emotion bearing) text has a good communicative and psychological value. for example to determine sentiments. However, it is unclear how exactly it works in CQA media like SO. Yet, seemingly, low-level users in SO like to exclaim more than average users. Conversely, within the Zooniverse content, that could be taken as a good check point whether users have made close observations of the images. Such category comprises the *exclamation mark* and the interjection *INTJ*.

Focus construction: represents users' effort in *focus constructions* in their text. That might be evident from some syntactic patterns present in parsed text. For instance, the syntactic tag *SINV* denotes users intentionally invert their sentence construction, probably to give more emphasis for the syntactic units intended to appear at the front. In comparison, such kind of linguistic style is more observed in the SO users than the Zooniverse counter part.

Social interaction: shows the *social* aspect of expertise. That is a quite important dimension of expertise, particularly in social media. Considering users' connection helps measure the quality of their participation and communication with others. According to some psycholinguistic analysis [1, 18], the occurrence of personal pronouns in one's text suggest his/her social interactions with others in a network.

Looking into our syntactic information, moods of verbs indicating third personal singular as well as general personal pronouns tags, *VBZ* and *PRP*, respectively, could be good candidates to form the *social interaction* category of expertise. In addition to that, social media symbols (e.g., *hash tags (#)*, *at-symbols* (a)) grouped together and added into the *social-interaction* expertise category. In practice, the presence of such symbols in one's social media text could indicate the user is associating and calling attention of others. In addition, combinations of possible *PRP* variants could define an *authoritativeness* aspect of expertise [18], but they occur much less in either domain.

Complexity: addresses either complex syntactic structures or unclear grammatical constructions as a result some piece of text get labeled with *X* by syntactic parsers. Because of that, *complexity* might confuse with *in-completeness*; also some tags appear to belong to multiple expertise categories.

Latent Topic Representation

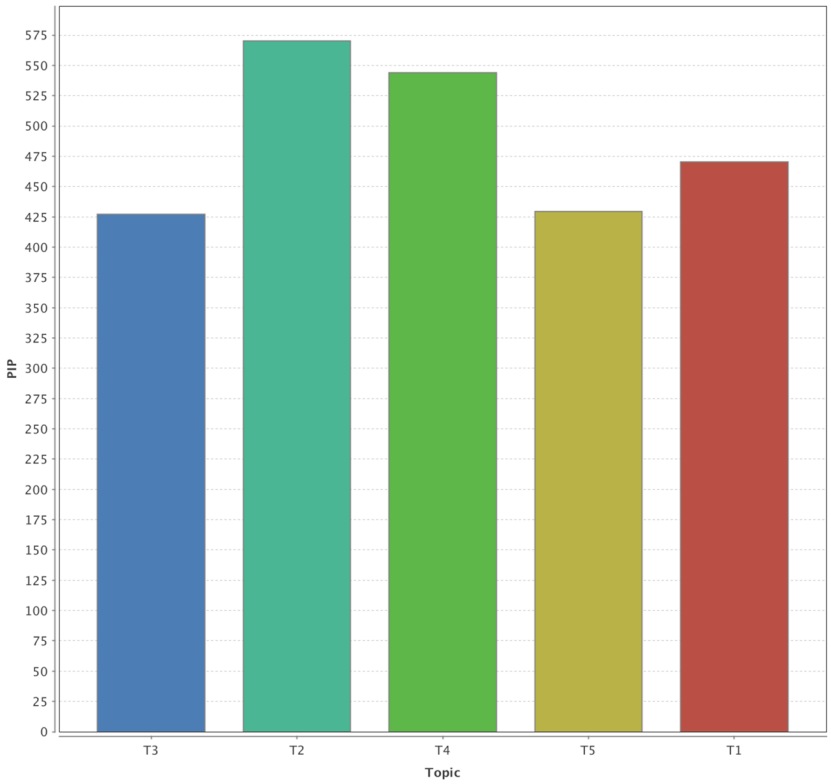
Looking both the SE and Zooniverse data, they are very much topic oriented. There are about 178 and 92 central topics being discussed by the crowd of the SE network and Zooniverse, respectively. To meet such topic-oriented setting, a quite specialized text representation method is more important than the generic bag-of-words model.

Apparently, experts in specific areas are likely inclined to frequently using more technically focused as well as valid content terms (right words), than non-expert crowds. Nevertheless, unlike the bag-of-words approach, critically selecting well-representative topic words (or referred as *expert words*), in advance, which fit the intended central topics, is always difficult. In addition, it is quite time taking unless we are very much familiar at an expert level with particular domains. Probably, having well-annotated dictionaries which are tailored precisely for a specific forum and, can act as *computational lexicons*, might be an alternative remedy, though the chance of getting such dictionaries is quite rare or, even impossible. In such condition, the bag-of-words (BoW) representation is quite a plausible choice, though it has potential downsides. It is *less-focused in terms of topics as it includes all sort of content and function words* and *computationally expensive*. In addition, BoW might contain hundreds of thousands of terms opposed to topic-oriented models.

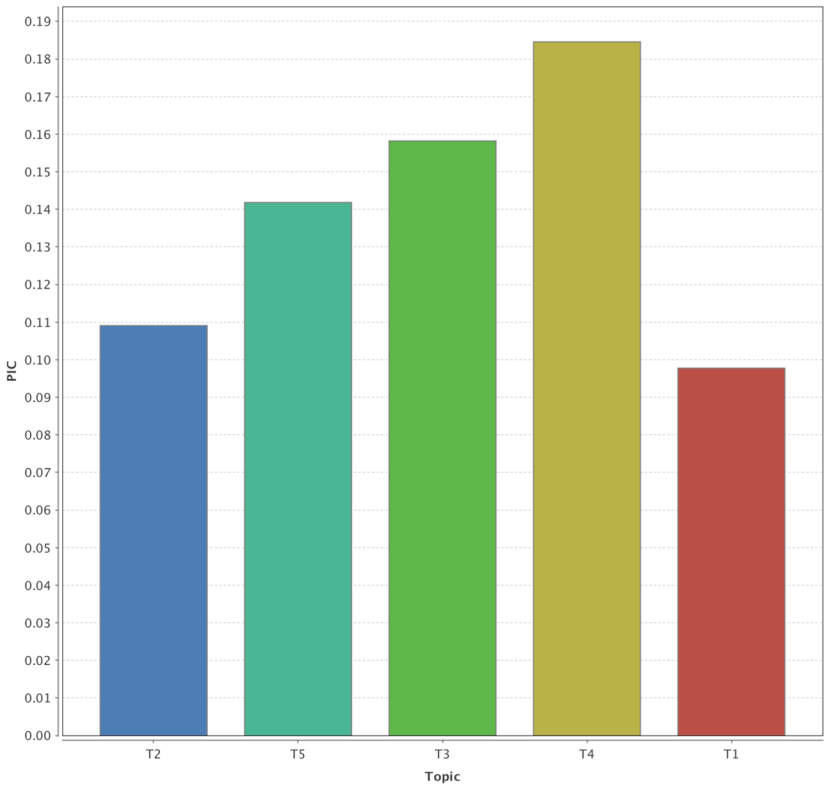
To avoid missing any important information about *function words* as a result of focusing on content words bearing some meanings and topics, their usage is captured during the syntactic representation. Incorporating, both types of information into the composition of our linguistic features, eventually somehow succeeds capturing *what* users basically *state in their content words* and *how* they say through their *function words*. Since technically the most relevant terms that could form expertise knowledge are unknown or (not identified) in prior, the BoW model tends to include all types of terms irrespective of their practical significance. That obviously affects the sharpness of the model for effectively making distinction between users whose text is quite focused (bearing domain knowledge) and generic. Once models got built on optimally identified representative terms, detecting users' expertise based on their topical information gets easier.

In this study, topic words have been extracted from both domains by using latent topic modeling [29, 30]. Initially, the number of topics along with the number of topic words in each domain have been determined in such a away that the size of the topic annotation corresponds to the size of the syntactic annotation. Such correspondence is assumed to make the later comparisons (with syntactic annotations) more sound. Thus, 65 latent topic words (5 topics * 13 topic words) have been extracted. In the later stages, the effect of varying the size of topic words has been analyzed as well. In contrast, the size as well as the average sentence length of Zooniverse is much less than SO, thus considering only topic words has been very important to handle and deal with the shortness of text. Moreover, as topic words tend to spread all over the text than generic words, they have the potential to capture and represent the central thought of the entire text.

To clearly view the variation of expert levels across topics, user have been grouped by their latent topics. In addition, the distribution of users over topics is shown in Figs. 5b and 6b for StackOverflow and Zooniverse, respectively. The largest portion of the distribution within StackOverflow is covered by **T4** (topic 4), and in Zooniverse by **T5**. The mean of both PIC and PIP scores of the users under similar topics have been computed for plotting their respective charts. That are shown in Figs. 5a and 6a for StackOverflow and Zooniverse, respectively. Figure 3 also illustrates the standard deviation of PIP and PIC scores of users discussing similar topics. The pair of topics and their associated mean scores allow to observe the link between expert level variations across topics. Differences among users grouped under similar topics. That is in terms of expertise levels have been presented (Fig. 3a, b) with a pair of topics versus standard deviation. Moreover, the presented statistical information (users distributions, mean and standard deviation) give important insights how users under similar ranges of PIP and PIP tend to discuss on a similar topic (Fig. 4). For instance, in Fig. 5a, users with highest PIC scores are likely focusing on **T1** and **T5** than any other topic. On the other hand, **T2** seems to be preferred by low-level users while **T3** and **T4** by medium-level users. Likewise, Fig. 6a shows high competent users fall under **T2** and **T4** while least competent ones emphasize on **T3** and **T5**.



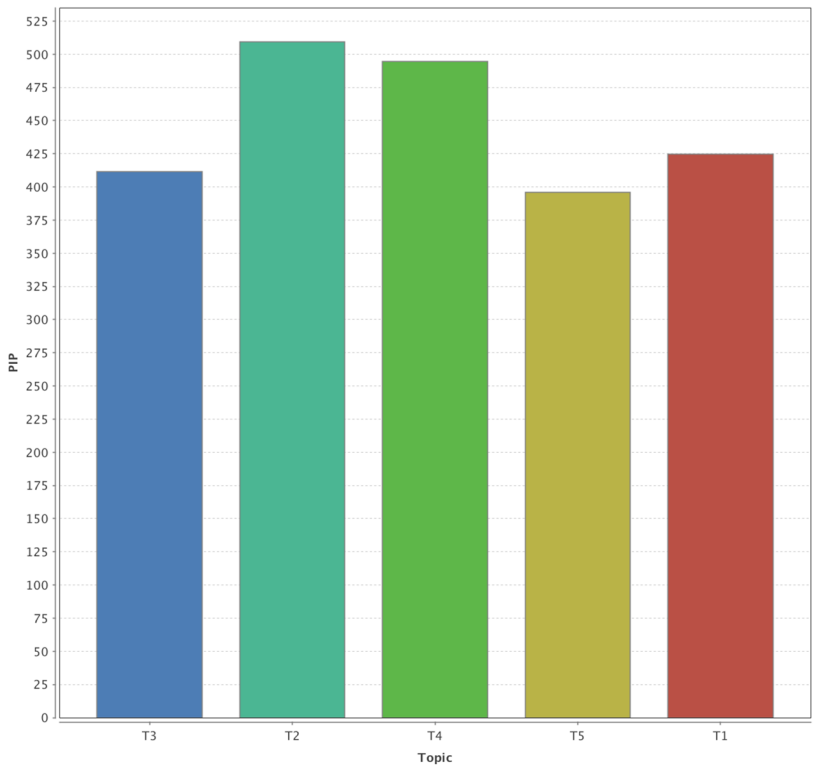
(a) Topics Versus Standard Deviation PIP Scores



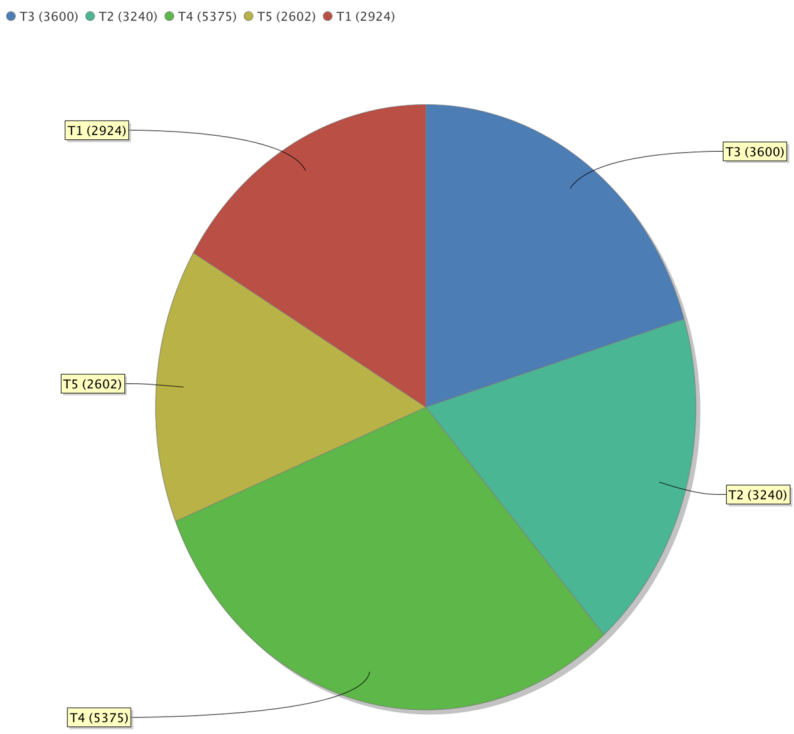
(b) Topics Versus Standard Deviation PIC Scores

Fig. 3 Variations of PIP and PIC scores of users within similar topics

Fig. 4 Grouping StackOverflow users on the basis of their topics and PIP scores

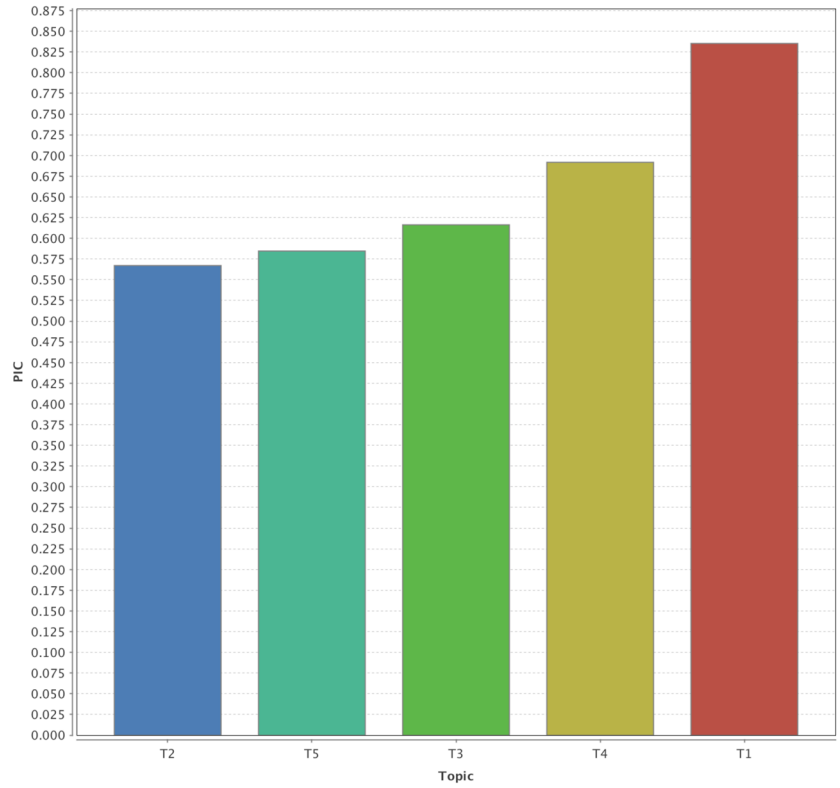


(a) Topics Versus Mean PIP Scores, *showing PIP scores rendering expertise levels vary as topics vary*

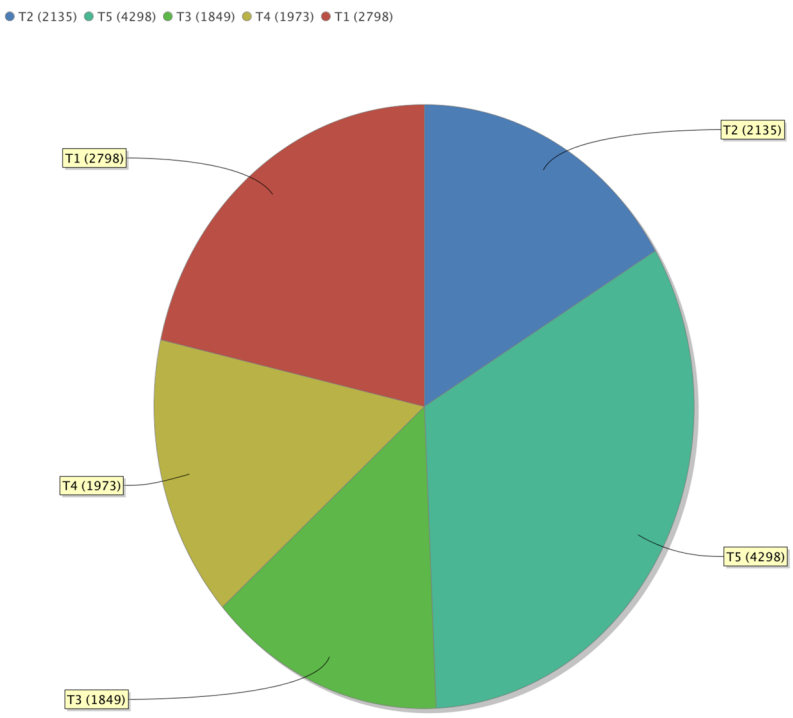


(b) Topics Distributions over the Frequency of SE Users

Fig. 5 Grouping Zooniverse users on the basis of their topics and PIC scores



(a) Topics Versus Mean PIC Scores, *showing PIC scores rendering expertise levels vary as topics vary*



(b) Topics Distributions over the Frequency of SEGA Users

Following latent topic extraction, three variants of topic word representations have been applied. The first one is *a unigram text representation*, where each users' document (a collection text messages posted by each user) is characterized by a feature vector of *unigram topic words (TW)*, or $\vec{d} = (\mathrm{TW}_1, \ldots, \mathrm{TW}_n)$, where n is a size of the unigram set, in our case, i.e., 65. The *a unigram text representation* has been used to build baseline models.

Second, *topic bigram and trigram representation*, where a sequence of two and three topic words, respectively, has been used to describe users' documents. Or their corresponding feature vectors could be like $\vec{d} = (\mathrm{TW}_1\mathrm{TW}_2, \ldots, \mathrm{TW}_{n-1}\mathrm{TW}_n)$ for the bigram and $\vec{d} = (\mathrm{TW}_1\mathrm{TW}_2\mathrm{TW}_3, \ldots, \mathrm{TW}_{n-2}\mathrm{TW}_{n-1}\mathrm{TW}_n)$, for the trigram representation. Bigrams and trigrams have been used as additional linguistic annotations to enrich the unigram baseline models with contextual and sequential information. Third, a *character bigram and trigram representation* [31] is applied to capture characters' sequential information [32, 33].

Expertise Modeling with Multiple Linear Regression

At this stage, gold-standard data have been prepared for training, validation and optimization, and evaluations, in the form $\vec{v} = (\mathrm{LAFe}_1, \ldots, \mathrm{LAFe}_n)$. $(\mathrm{LAFe}_1, \ldots, \mathrm{LAFe}_{n-1})$ and $LAFe_n$ denote linguistic, and possibly non-linguistic annotations and proficiency scores, respectively. Both types of proficiency, PIP ($\mathrm{pS}_{\mathrm{PIP}}$) and PIC ($\mathrm{pS}_{\mathrm{PIC}}$) are quantified with continuous numeric scores, ranging $[1, \infty)$ and [0, 1], respectively. Having such numeric scores serving as labels (target variables), and multiple independent variables acting as predictors, multiple linear regression models have been built. Multiple linear regression is chosen as we are dealing with a regression problem and attempt to model continuous target variables (PIP and PIC). Basically, the multiple linear regression models' equation [34, 35] include slope coefficients $(\beta_1, \ldots, \beta_n)$ for each linguistic annotations feature and bias (β_0):

$$\mathrm{pS}_{\mathrm{PIP}} = \beta_0 + \beta_1\mathrm{LAFe}_1 + \beta_2\mathrm{LAFe}_2 +, \ldots, +\beta_n\mathrm{LAFe}_n. \quad (2)$$

Training and Validation

Baseline and *enhanced* versions of expertise models (about 20 different models built) have been trained. The models are iteratively trained on the linear combination of various linguistic annotations discussed above. The *adaptability* of the baseline models allows to add and learn any types of linguistic information in addition to its core composition of features intended to be experimented. The baseline model trained on *unigram topic words*. In addition, the enhanced models on the additional richer linguistic annotations intended to improve baseline models' performance. Basically, the added annotations include *syntactic and punctuation annotations*, *semantic and character sets annotations*, *word and character n-grams (n = 2 and 3) annotations*.

Prior to the actual models' evaluations phase, we made sure that each model is statistically fit. Thus, *validation and significance tests* have been carried out as part of the training of each model. Running such tests on the models help ensure their validity and reliability. In addition, that ultimately guarantees the learned models are *statistically significant*. Unlike the cross-validation approach, which makes uses each example of the entire corpus for both training and validation, we apply split-validation tests on independent development test sets separated from the training and the evaluation sets. Setting the validation sets apart, in contrast, ensures the reliability of the models better, as it avoids using the same instances of examples twice or more throughout the models building life cycles. The *null-hypothesis* ((H_0)) [36] test, which is a standard test of significance for regression models, has been applied using development sets, with a threshold alpha value (p-value) of 0.05, for ensuring whether the slopes of the models are different from zero. That implies that the slopes associated with the selected linguistic annotations could either positively or negatively describe the target expertise scores, significantly:

$$\mathrm{pS}_{\mathrm{PIC}} = \beta_0 + \beta_1\mathrm{LAFe}_1 + \beta_2\mathrm{LAFe}_2 +, \ldots, +\beta_n\mathrm{LAFe}_n. \quad (3)$$

Model Evaluation

Following the validation phase, the prediction performance of each model has been evaluated on seven different test sets. Such test sets are collected from a wide variety of crowdsourcing domains. The evaluation metrics are R^2 *(R-squared) and RMSE (root mean squared error)* [37]). The former (aka *the coefficient of determination*) measures how well the learned models (the selected linguistic features) describe the variation in PIP and PIC scores, while the later (aka a *residue measure*) measures prediction errors.

The test sets include in-domains (and related domains), where the domain of the evaluation and the training data is exactly the same or very much related. Out-domains test sets are completely independent (or different) from the training sets. In addition, there are two in-domain datasets selected, namely SO and Galaxy Zoo and Snapshot Serengeti. They belong to the same domains as the training set. In addition, two related domain test sets: Server Faults and Suser, and three out-domain datasets, English, Math and Ubuntu have been in the evaluation. The aim of considering such

diversity of domains, is to get the insight that to what extent the learned models could somehow generalize the crowdsourcing forums. That is also important to build more assertive claims on them. In the evaluations, the two most critical aspects of any regression model quality have been estimated. The potential of the models to effectively explain the variation in the expertise scores (PIP or PIC scores) and the overall performance to reduce prediction errors, via the coefficient of determination, i.e., $(R)^2$ [35], and RMSE [34], respectively.

Since the diversity is not only in terms of domains, it also includes, the range of the target variable, i.e., expertise scores as well as the size of the test sets. To best address the diversity, the performance measurement, particularly the predictions errors aspect, has been geared towards considering such main differences among the test sets. Therefore, two normalizing techniques [38]) have been applied, namely *RMSE mean normalization* and *RMSE range normalization*.

Results and Discussion

In this section, we provide qualitative and quantitative analyses. In addition, comparisons of the overall evaluation results yielded by the learned expertise models are provided. For the former case, we provide a summary of results along with comprehensive interpretations. For the later case, we support the analysis with ordered lists of various linguistic annotation sets from both sides (StackOverflow and Zooniverse).

We also illustrate how the baseline models have been enhanced by incrementally adding more linguistic annotations on top of them. The selected linguistic annotations include syntactic, punctuation, word bigrams and trigrams, and character bigrams and trigrams. Practical significance of these models have been also discussed briefly.

Baseline Models

The baseline models are built on *unigram topic words* for both forums and their evaluation is summarized in Table 3. In addition, important commonalities between the two forums' models have been identified, which in turn lead to the fact that the models trained on the selected CSFs could be adaptable to other related forums. Concerning R-squared, while the best model (i.e., the PIP model evaluated on the Server Fault category) scores 0.7940, the poorest model (i.e., the PIC model evaluated on the unified Galaxy and Snapshot Serengeti datasets) scores 0.3100. The results illustrate the selected topic words are able to explain approximately up to 79% variations in PIP.

Compared to regression models built particularly in the area of CSFs [5, 18, 39], the result found in this study is quite promising, and has some limitations as well. For instance,

Table 3 Unigram topic words baseline models' evaluation results

Forum	Category	RMSE-m	RMSE-r	R^2
SE[a]	SO[b]	0.5114	0.0557	0.7290
	UB	1.3192	0.0608	0.5740
	SE	0.4728	0.0495	**0.7940**
	SU	0.5393	0.0470	0.7830
	EN	0.9365	**0.0263**	0.6850
	MA	0.7372	0.0529	0.7120
ZO[c]	GASS[d]	**0.2200**	0.1400	0.3100

[a]StackExchange

[b]StackOverflow

[c]Zooniverse

[d]Galaxy Zoo and Snapshot Serengeti

Highest scores across each metric have been provided in bold text

the authors in Ref. [18] found 0.40 and 0.07 R-squared, on the linear regression models trained on linguistic and non-linguistic meta-features to predict answer and question quality, respectively. Predicting question quality has been also studied in Ref. [39], using question-related attributes, and the authors reported 0.19 R-squared. On the other hand, our topic-based approach shows significant improvement over the BoW approach used in Ref. [5], in which the highest R-squared score provided by the trained models is 0.74. That means, the method developed in this study has pushed the R-squared value from 74 to 79%. Nevertheless, some of the compared models outperform the models built in this study, in terms of RMSE. For instance, the model built in Ref. [39] has lower errors (i.e., 0.19) than ours (i.e., 0.2200).

Such a significant shift of R-squared (from 74 to 79%) implies that how applying topic modeling effectively improves the performance of expertise models over common bag-of-words models. There is no standard procedure for guiding the choice of optimal words that best fit expertise, in prior to relevant (preliminary) experiments. Thus, most generic text-based models are tend to be built on a BoW model that often contains several thousands of terms extracted from training corpora. That has a number of downsides besides poor model quality, for instance posing a huge overhead of computation resources.

Domain-wise, comparatively the StackOverflow model seems to be more effective in predicting user expertise than the Zooniverse model. There are a couple of possible reasons. First, PIP scores always have a direct relationship with the textual content, particularly answers content contributed by StackExchange users as such points are awarded based on the perceived quality of the answers. Nevertheless, the PIC points are exclusively estimated by considering only the number of classifications made by Zooniverse users. Therefore, the computation of PIC is independent from their corresponding text. Second, the StackOverflow model

Table 4 Baseline and enhanced models' evaluation results

Model	RMSE-m	RMSE-r	R^2	MAX	MIN	Min-m	Min-r
Baseline	0.6766	0.0618	0.6553	**0.7940**	0.3100	0.2200	0.0263
Baseline+S[a]	0.6958	0.0631	0.6561	0.7880	0.3210	0.2200	0.0270
Baseline+P[b]	0.6757	0.0617	0.6539	0.7930	0.3090	0.2200	0.0262
Baseline+S+P	**0.6685**	**0.0601**	0.6643	0.7930	**0.3900**	**0.2000**	**0.0261**
Baseline+B[c]	0.6782	0.0619	0.6653	0.7920	**0.3900**	0.2200	0.0264
Baseline+T[d]	0.7282	0.0660	0.6089	0.7450	0.3563	0.2200	0.0274
Baseline+CBT[e]	0.6770	0.0617	**0.6663**	0.7890	**0.3900**	0.2200	0.0267

[a]Syntactic annotation

[b]Punctuation mark annotation

[c]Bigrams

[d]Trigrams

[e]Character bigrams and trigrams

Highest scores across each metric have been provided in bold text

is purely trained on an in-domain dataset, on the other hand, the Zooniverse model has been built on the mixed datasets. Thus, the chance of mis-recognizing unseen inputs is a bit higher due to relying on half knowledge of either category. Moreover, there is also a notable training size difference (in terms of number of users and chat messages, average sentence length) between them.

Application-wise, we primarily aim to build models that learn expertise from linguistic features and improve the quality of CSFs. Those topic words identified as most relevant to infer expertise, however, could be used to address other important aspects for many other interesting purposes. For example, getting users answering questions friendly and politely (in their expressions) is a concern in SO and implicitly connected with their reputation. For instance, the StackOverflow question-answering's guideline[9] explicitly mentions and suggests some selected words to be taken into account during answering questions to ensure certain moral standards. Since for some users (askers) the way an answer is written/presented (typically wording and delivery) for a particular question, perhaps is as equally important as the answer itself. Nevertheless, it is quite hard for CQAs to inspect users' content and fully address such concern, without prior modeling of users' word usage patterns. That might in turn lead to failure to achieve the goal of having trustworthy and morally responsible users. As also is evident in some studies [40, 41] on how moral behavior quality reflected in answering questions in social media affects readers.

Error Analysis and Improving the Baseline Models

We provide illustrations of error analyses and the relative importance of iteratively added linguistic annotations for enhancing the baseline models. For better insights, detailed demonstrations for each annotation type has been provided in Tables 4 and 5.

We measure and analyze the significance of the improvements from various perspectives based on major statistical parameters as criteria. Many of these perspectives prove that integrating those annotation features make important contributions. Nevertheless, exceptions have been identified as well. The impact of adding punctuation, in contrast, is not as great as others. Therefore, we rather discuss with its linear combination with syntactic annotations.

Model Enhancements by Adding Syntactic Annotations

R-squared-wise, syntactic annotations enhanced the baseline models of both expertise types. Particularly, the PIC model has been improved by 3.43%. These annotation features have been added with the expectation that the syntactic structures to be learned from the in-domain datasets. That could help detect typical syntax use in out-domain (related) datasets. Because of that, their impacts have been, particularly assessed and emphasized on out-domain

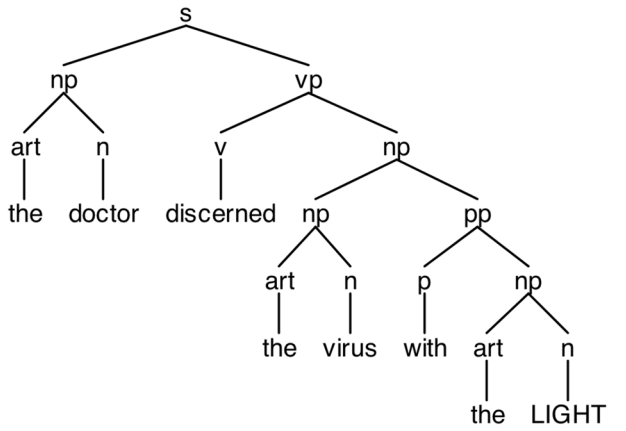
Table 5 Syntax + unigram topic words or enhanced (with *syntactic annotations*) models' evaluation results

Forum	Cat.[a]	RMSE-m	RMSE-r	R^2
SE	SO	0.5702	0.0621	0.6980
	UB	1.3993	0.0645	0.6010
	SE	0.4961	0.0520	**0.7880**
	SU	0.5351	0.0467	0.7800
	EN	0.9594	**0.0270**	0.7060
	MA	0.6903	0.0496	0.6990
ZO	GASS	**0.2200**	0.1400	0.3210

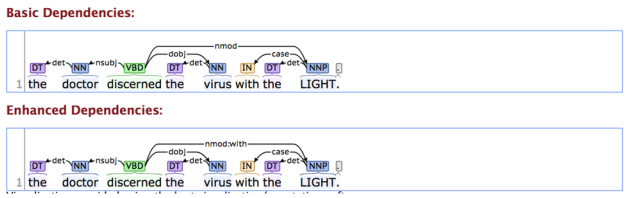
[a]Category

Highest scores across each metric have been provided in bold text

[9] https://stackoverflow.com/help/how-to-answer.



(a) A Constituency-based Parse Tree



(b) A Dependeny-based Parse Tree

Fig. 6 Syntactic and semantic representations of the sentence *The doctor discerned the virus with the LIGHT*

datasets. In this regard, interesting results has been found. For instance, while the majority of the test case results on out-domain datasets have been improved. The maximum shift of R-squared has been caused as a result of adding syntactic annotations.

In the best case scenario of out-domain test sets, the English dataset has given the best result, regarding both versions of RMSE and R-squared. Observing the English test set being well recognized by the improved model trained on a fairly different dataset (i.e., the SO dataset). That implies the distinctive role of syntactic annotations is to build quite general models. That also could serve *inter-domain forums*. That means, inferring the expertise of English forum users has been possible by capturing syntactic structures than actual content. However, the expertise of StackOverflows' users is different.

Relative importance of the syntactic features are analyzed. In addition, the analysis shows that part-of-speech tags (word-level constituents) (e.g., *LS, UH, and RP*) are dominant predictors. They are also negatively correlated with expertise scores. That supposedly implies that low-level experts seem to like making lists of items and express their opinion with interjections, more than high-level experts. There are also observable differences between expert groups in named entities use, denoted by *NNP* and *NNPS*. Least competent users frequently use the syntactic categories *NNP* pretty much frequently than *NNPS*. But top-level users do the other way round.

The other important aspect of the improved models is how it penalizes users, for using *foreign words* in their expressions. For instance, the PIC model cuts off PIC scores from those users who use non-English terms. That naturally makes sense, as mixing *words containing non-English characters* in English-based forums might confuse the crowd. On the other hand, there are exceptional cases, where names of objects. For example, within the Snapshot Serengeti forum, some species could have typical foreign names due to their origin. Thus, handling such exceptions for example through an additional NER (*named entities recognition*) task might improve the accuracy of the model than only relying on *NE-like syntactic* (*e.g.,* ***NNP, NNPS***) *tags*.

Model Enhancements by Joint Syntactic and Punctuation Annotations

We dedicate this particular sub-section for error analysis to identify core causes that make the prediction wrong. In addition, we suggest possible strategies to reduce prediction errors. From any other linguistic annotation, the linear combination of syntax and punctuation played a great role of reducing prediction errors of the baseline models. Such composition resulted the least RMSE average value, compared to all added annotations. On average, the errors of the baseline models' declined approximately by 9%, as a result.

For the error analysis, we filter test cases, that could establish a dichotomy between well predicted versus poorly predicted. These two marginal groups are sampled from seven different categories with the most significant RMSE as well as the exact opposite of them, from all evaluation sets. That means, from each category two groups of users have been set aside. These users could serve as representatives for (almost) 'perfectly predicted' (referred in our discussion as first group) and 'poorly predicted' (second group) examples. Subsequently, we explain possible reasons from the perspectives of the linguistic features identified as dominant in the structure of the mixed syntactic and punctuation annotations based models. The error analysis along with descriptions is, thus, based on certain criteria. These criteria involve *actual and predicted expertise scores*, *the identified groups of users*, *models' structures* and *the most prominent predictors from both types of annotations*.

We consider the PIP predictions errors resulted from related categories test sets, particularly those with highest mis-recognized users. Their parse trees are enriched by diverse syntactic categories, and larger in quantity as well. For instance, the syntactic category **WP (Wh-pronoun)** does not occur in the second group at all. In addition, it is observed in 80% of the first group. Similarly, the associated tag **WHNP (Wh-noun phrase)** only occurs once in the

second group. Moreover, it has exactly the same coverage as **WP** in the first group. Lacking such syntactic categories in the second group has led the model to the resulted prediction errors. In contrast, they happen to favor almost the accurate recognition of the first group.

Nevertheless, alternative same expressions without explicit use of these categories might be present in the linguistic construction of the second group, though hard to be noticed by the models. For instance, in the piece of text *I know* ***what*** *the class of the galaxy, when I look at the pattern of the surrounding stars*. That has been posted by the first group users. Such type of post leads a better prediction due to the occurrence of the **Wh-pronoun** *what*. Quite an equivalent statement can be made without explicitly using or even by cutting off the **Wh-pronoun** *what*. Nevertheless, due to the fact that the syntactic-based models are trained to be *more sensitive to symbolic information* than *actual expressions*. That very similar statement made by the second group gets mis-recognized. Thus, a possible practical strategy to address this linguistic phenomenon and reduce the errors is to completely identify such types of expressions. In addition, treat them under similar syntactic categories.

Domain-wise, most of the PIC prediction errors emanate from least competent users. Looking into the size of their text, it is approximately 4 times smaller than users' text size posted by correctly predicted users. Probably, considering special features inherently characterizing very short text, microblog posts or that generally best suit for writing styles of these particular type of users reduces the prediction errors [42–44]. On the contrary, looking punctuation usage patterns, for example, period marks' usage of the best recognized users, is quite less frequent (approximately twice less) than the mis-recognized users. Yet, the former group has a pretty much uniform distribution of periods. In addition, the sensitivity of the models to exclamation marks has been to a possible reason for either correct prediction. Or failures to detect those users who rarely exclaim explicitly with exclamation marks, might be the cause as well. This feature exceptionally occur, particularly in the textual content of Zooniverse forums. That is from any other punctuation mark, because of spectacular images of galaxies or wildlife shown to users. That makes the logic behind the sensitivity of the PIC models to emotional expressions clear. Nevertheless, it is unclear in case of the PIP models. A partial remedy for this particular issue is perhaps to develop or integrate sentiment analysis tools. Such tools are able to detect deep linguistic phenomena (e.g., sarcastic expressions, implicitly polarized statements) likely buried within ironic expressions.

The punctuation mark **"."** and other complex punctuation patterns cause prediction errors remain unclear as well. Thus, further analysis could reveal the potential reasons. It gets a bit more complicated when it comes to the PIP models, due to the fact that punctuation marks have multiple different senses in StackExchange. That is because they get mixed with character sets occurring in code snippets. For instance, in natural text, the punctuation mark "!" is typically used to exclaim, whereas in programming (or in scripting languages) it serves as an operator (e.g., the **not** operator). Thus, feeding the expertise models this mark without sufficient context might confuse the learning. In addition, that eventually result *model perplexity*, which is a quite typical concern in language modeling [45, 46]. Thus, making clear distinction between punctuation marks and character sets during the linguistic annotation phase likely reduces the perplexity (disambiguate the confusion). As a result, the models are likely to get shrewd.

The other interesting fact observed from both groups of users is their *questioning habit*. There seem to have quite a distinctive trend between them. While the first group asks many questions, their counter part remain reserved from inquiring. That implies the model has well learned treating users posting *interrogative statements* than *simple declarative expressions*.

To resolve this, either increasing the proportion of examples bearing declarative statements in the training set. Or that is to introduce new features into the model that particularly target and heavily handle this condition. The former strategy helps avoid data imbalance (skewness) and improves the learning out of sufficient observations containing declarative statements. On the other hand, to mine the potential for effectively identifying the new distinctive features, the later strategy requires to closely look at the dynamics of that particular group of users in connection with their PIC scores.

It is also important to note that the underlying problem for most of the prediction errors lays under the linguistic annotations techniques and tools used in this study. That might include syntactic parsing errors along with the associated part-of-speech tagging flaws and the inaccuracies of other downstream annotators (e.g., named entity recognition). In addition, expertise scores estimation errors, eventually lead the subsequent regression errors.

Since, fundamentally, syntactic parsers get trained on manually annotated (by humans judges) corpora of parse trees based on tagging conventions enforced in certain annotation systems. These parsers make probabilistic decisions which is quite error-prone. Unambiguously, specifying boundaries of certain syntactic categories is not always easy, as well as subjective. For instance, **NAC** is one of the significant syntactic categories, which serve as a scope marker of modifiers within a noun phrase. Any error occurring as human annotators tagging input text (training data on which the resulting parser models get trained on), obviously propagates to the learned syntactic-based models built in this study. Thus, using more accurate annotation tools perhaps would improve the results.

Analyzing the *actual expertise scores* against their corresponding *predicted values*, least competent users are more wrongly detected as 'high competent' than the other way round. That means, as evident from dominant features, such group of users satisfies the *linguistic qualification requirements*, imposed by the regression models, and they got predicted as 'high competent'. Yet, they belong to the 'low level' class as their *actual expertise either* ***answering questions*** *or* ***classifying images***, is not that great. The underlying possible reason for the mis-recognition might be linked with the PIC/PIP scores estimation method. Thus, improving the expertise estimation algorithms to balance the *practical aspects or the* ***deed*** *part* with *linguistic qualifications or the* ***word*** *part*, might reduce the errors.

Model Enhancements by Adding Bigrams and Trigrams

Adding bigrams and trigrams has slightly improved the performance of the baseline models. Particularly, while bigrams cause the mean R-squared to shift from 0.31 to 0.39, which approximately a 20.51% improvement over the baseline, in the evaluation on the Zooniverse test set, (as shown in Table 4). On the contrary, adding trigrams has been more effective than bigrams as they provide larger contextual information [47, 48]. In most cases, natural language modeling tasks, considering sequential and contextual information yields good results. But there are some exceptional cases where its impact gets limited [49]. We explain that with simple examples observed in our study in the next paragraph.

As opposed to common *N-grams* in *bag-of-words*, terms in topic *N-grams* do not often occur very much neither get arranged (co-occurred contiguously). That is because non-topics words do in natural text. For instance, from the piece of text taken from a moderately competing Zooniverse user **"Green light peeping through on the center left, 4 object in a line to the left of galaxy, oval shaped galaxy."** the bigrams ***green-light, light-peeping, peeping-through*** and trigrams ***green-light-peeping, oval-shaped-galaxy*** could be generated only if the conventional *BoW N-grams* model has been applied. In case of *topic N-grams*, however, only the bigram ***green light*** and the trigram ***oval-shaped-galaxy*** are considered, as governed by the learned *topic words (e.g., light, galaxy, green, shape, oval)*. That means, the excluded *N-grams* are unlikely to occur in the text representation of topic words arrangement. As a result of that, the impact of the learned *topic N-grams* models has been so great. Perhaps, hybridizing the traditional *N-grams* with *topic N-grams* possibly produces better results, although at the cost of losing (deviating from main terms) focus.

Model Enhancements by Joint Character Bigrams–Trigrams

Evidently, the largest shift of mean R-squared has been scored by adding joint character bi(and tri) grams. It also improves the PIC model with the same magnitude as *topic N-grams*. Among other datasets, adding character n-grams has dramatically reduced the errors of the Maths category. A partial reason for that might be the extracted character bigrams and trigrams happened to match ***few-letters*** *alpha numeric notations. That represent variables and quantities*, in mathematical expressions and favored for the reduced RMSE.

That further sparks an interest in how the influence of *topic N-grams* on other out-domain test sets (i.e., English and Ubuntu). The impact is also contrasted with in-domain (i.e., Galaxy, Snapshot Serengeti and StackOverflow) and related domain test sets (i.e., Server Fault, Super User). The evaluation results show on in-domain datasets, the mean enhancement gets approximately 3%, R-squared-wise. On the related domains, e.g., Super User gives a slight improvement of R-squared (from 0.7830 to 0.7840). But the result declines on Server Fault from 0.7940 to 0.7890. On out-domain datasets, there is an average enhancement of R-squared from 0.6570 to 0.6637. Thus, the overall assessment of adding joint *character n-grams* on top of baseline models leads to the conclusion that morphological (character-level) aspects of text could be best dealt with variable length *character n-grams* [33, 50].

Further Improving the PIP Model

As noted in the results, the PIP models are likely to enhance and respond to the added linguistic features better than their PIC counterpart. Thus, moving further towards enhancing the PIP models might make more sense. To advance our understanding how other types of linguistic information influence the prediction of PIP scores, *semantic* and *character set information*, have been added on top of the *unigram topic model PIP model*. We discuss their relative effects in the next subsections.

Influences of Joint Syntactic–Semantic Annotations

By linearly combining semantic information together with the syntactic annotation, almost a *full enriched linguistic annotation (structure-wise)*, has been achieved. Given that both help capture structural information from text, semantic particularly focuses on *dependency structures* or *dependency relations between words*, while syntax emphasizes on *phrase structures*. To extract those dependency relations, the training textual corpora have been parsed with a *dependency parsing algorithm* [51, 52]. The parser generates the corresponding *dependency parse trees*. Subsequently, *head*

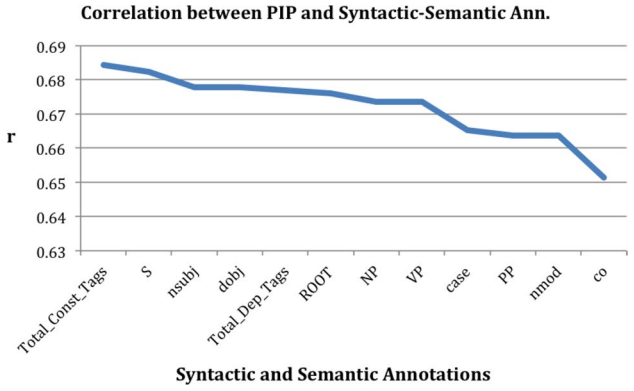


Fig. 7 The correlation between syntactic and semantic annotations and PIP

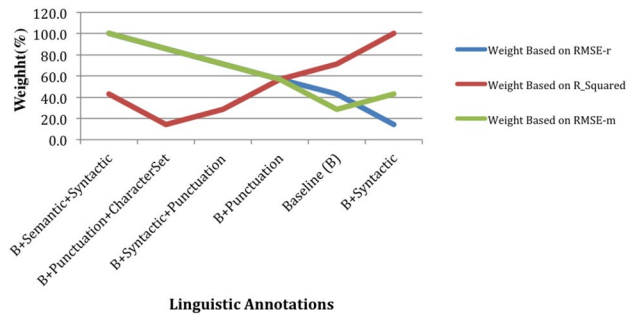


Fig. 8 Percentage of weights for baseline and added linguistic annotations, *depicting an ordered list of linguistic annotations ranked based on their average weights, computed from their R-squared and RMSE scores*

words, and other dependencies (e.g., *nsubj, case, dobj*) present in the generated parse trees, are extracted.

A simple example is given in Fig. 6 to illustrate main differences between syntactic and semantic representation. In this example, the sentence *The doctor discerned the virus with LIGHT.*, has been parsed with constituency and dependency parsers [26, 27, 53]. Within the resulting dependency tree (shown in Fig. 6b), *eight dependency relations* occur. These are det(doctor-2, The-1), root(ROOT-0, doctor-2), acl(doctor-2, discerned-3), det(virus-5, the-4), obj(discerned-3, virus-5), case(LIGHT-8, with-6), det(LIGHT-8, the-7) and obl(discerned-3, LIGHT-8). While the word **discerned** is the *head word*, all the remaining terms are *dependents* (Fig. 7).

Interestingly enough, the *joint semantic and syntactic annotation* play a consistent role of reducing errors. It has also improved R-squared better than any other combination of annotations (as shown in Chart **2**). How they are correlated with PIP scores has been provided in Fig. 8. The relative influence of all the considered combinations has been weighted and ranked, and illustrated in Chart **2**. The highest weight is assigned to joint semantic and syntactic, based on relative significance improvement regarding both R-squared and RMSE over the baseline model. Among the added semantic features, *compound:rt, exl, csubj* positively influence PIP scores. But *root, parataxis, det:redet* have a negative influence.

Moreover, all test sets show improvements, except the Maths set. Perhaps, adding semantic information on the numeric data that involves mathematical expressions, equations, algebraic notations and so on does not seem to affect the predictions (almost insensitive for the added semantic info). New approaches that directly deal with typical patterns of actual Maths content, such as NER might help for numeric data detection. However, comparing with results obtained in other related studies, for instance, Crossely et al. in Ref. [54], were able to find 30% R-squared, their linguistic features could explain only 30% variance in math performance as opposed to our result, i.e., approximately 70%. Noticing that, we used the Math data only for an evaluation purpose, perhaps better performance results could be achieved if it was used as training data.

Influences of Adding Character Set Annotations

In text-based environments, where *natural text* mixed with *programming constructs*, considering both punctuation marks and *character sets* helps capture the full picture of *symbols' usage patterns*. Moreover, that helps avoid being conjectured on partial punctuation information. Applying the full version of punctuation has particularly improved the test result of StackOverflow, and slightly Math. Unfortunately, it has no any effect on English. That quite fits with the expectation that enriching the baseline model with character set annotations (character encoding information). That primarily benefits those categories (domains) containing code snippets written in various programming languages, i.e., StackOverflow. In addition, anticipated to propagate towards and affect Math's results. That is because some of them, particularly, the significant ones act as mathematical notations as well. On the other hand, some of the added symbols (e.g., curly brackets, a dollar signs) are not expected to appear very much in some domains. The best example is English. For instance, in the English dataset, *open* and *closed curly brackets* occur only 150 times. But in StackOverflow, they appeared about 34,000 times. For some reasons, however, among the significant symbols, an **open curly bracket** together with **closed square bracket** negatively influence the PIP prediction the most. However, characters (e.g., tilda, reminder, parenthesis) have a positive effect.

Considering typical good programming practices, suggests why using too many instances of an *open curly bracket* as well as a *closed square bracket* leads to the reduction

of PIP scores. Either overloading code snippets with several symbols generally makes other users less interested, eventually vote-down their answers, or leaving an opened (applies for the closed bracket too) curly bracket unclosed (un-delimited) is confusing users asking questions. That possibly affects the perceived expertise of answerers. That also clearly exhibits the interaction between *content quality characterized by selected punctuation* and *non-linguistic features (e.g., number of votes)*.

Next to *the majority linguistic annotations*, the selected *non-linguistic features* have also been important to some extent to define the expertise models. Nevertheless, that is not in the case of the PIC models. However, the main focus and the scope of this study is to investigate particularly the **impact of linguistic constructions** on the **detection of users' expertise**. The linguistic constructions in the natural sense, rather than the informal sense (programming/scripting languages' sense). But, it would be very interesting to go beyond from the natural, completely identify and learn typical code patterns practiced in programming/scripting environments and eventually related with users' expertise.

Further looking into the actual datasets of other domains, reveals other exceptionally occurring symbols. These symbols provide interesting information on the distributions of punctuation across domains. They potentially raise many philosophical questions, as strangely enough, they are still appearing colloquially in plain text of the English corpus. Knowing this information in advance, and other related linguistic trends, also, somehow helps to effectively identify linguistic features during model planning. Unfortunately, as most existing analytics particularly on question-answering forums apply meta-features instead. Therefore, discovering commonly followed linguistic patterns by CQA users requires separate research efforts that take the uniqueness of the data into account.

Influences of Adding More Topic Words

In addition to the baseline models described above, another two models have been trained. The training is done by incrementally doubling the pair of *the number of topics* and *the associated topic words*. The idea is to observe performance changes by stretching the size of the set of topic words by which the original baseline model has been trained on. The largest baseline model is built on *2340 (30 topics*78 words) unigram topic words*. In addition, the second model contains *1625 (25 topics*65 words) terms*.

The overall resulting evaluations show that the baseline model with least number of words outperforms the largest baseline model. Thus, keeping the size (in terms of number of topic words) small may help the model focus on very important words and seems to be an optimal choice. For example, some of the added topic words (e.g., *objective, property, and address*) are quite generic (as opposed to code-intensive terms (e.g., *int, void, return*) or names of programing languages as well as their associated keywords (e.g., *Java, PHP, function*), constructing the original model (e.g., *int, void, return*)), not uniquely identifying that particular domain, instead that cause the original model to loose its sharpness. But, it would be interesting that further in-depth analysis with additional topic words, surely offers a better insight and strengthens the conclusions.

Conclusions

In this study, we attempted to answer what defines typical **crowdsourced text expressions**. That is mainly in terms of *syntactic structures*. In addition, it is achieved through an extensive **characterization of major expert groups' linguistic constructions** and **joint syntax–punctuation annotations analysis**. **World's largest crowdsourcing forums**, namely Zooniverse and StackExchange, have been targeted, for the analysis.

As a result, six different **joint syntax–punctuation patterns**, have been identified. These patterns, potentially, allow to quantify and measure differences between expert groups in their *linguistic constructions styles*. Significantly observable differences across these expert groups have been discovered. The patterns also help identify 9 **text specific expertise dimensions** associated with linguistic qualities. Essentially, that could help establish a *standard linguistic-based framework*, to define and assess crowds' expertise within CSFs on the basis of their *linguistic constructions' quality*.

Latent topic modeling analysis also confirms the presence of certain differences between expert groups with respect topic words use. The analysis further reveals the relationship between users' competence and their associated topic preferences. More importantly, the **latent topic document representation** used in this study has shown quite significant impact over the common **bag-of-words representation**. A 5% improvement has been achieved in the prediction of expertise scores, through **latent topic annotation**, in comparison with the **bag-of-words approach**.

Other wide varieties of linguistic annotations: *syntactic and punctuation annotations*, *semantic and character sets annotations*, *word and character n-grams (n = 2 and 3) annotations*, have been extracted. The most significant *linguistic factors* which determine *expertise levels* of crowds have been identified. They have been weighted and ranked based on their impacts on the prediction of expertise scores, illustrated in Fig. 6. Their complete ordered list regarding R-squared includes **baseline topic unigram (B)+Syntactic+Semantic, B+Punctuation+ Character-Set, B+Syntactic+Punctuation, B+Punctuation, B** and **B+Syntactic**.

About 20 different **expertise models** have been built on the selected linguistic annotations. The *baseline* models, which are **unigram latent topic words based**, have been iteratively *enhanced* in a *two-stage process*. That is achieved by adding syntactic, semantic, word and character *n*-grams punctuation and character set annotations. That demonstrates the **extensibility** of the models to adapt other types of linguistic features.

Model validation and evaluation have been carried out on **8 crowdsourced corpora**: 2 of them are **in-domain test sets**, the remaining datasets are **related** and **out-domain**. The evaluation results on such quite heterogeneous collections of data guarantees the validity of the learned models across domains. The best model yields nearly, **0.7940 R-squared** and **0.0263** and **0.2200**, **RMSE**, range and mean normalized, respectively.

Major Insights and Future Work

Following quite extensive analysis of model evaluation results, interesting insights into computational links between proficiency and major linguistic patterns have been gained. The analysis revealed significantly a strong link between the quality writing styles and user proficiency. That implies high credibility of rating systems of forums could be achieved by deeply looking into users' writing styles than superficial meta-data cues.

In addition to improving rating systems, the identified joint syntax–punctuation patterns could also be used to enhance other NLP systems. For instance, authorship attribution, sentiment analysis and answer quality predictions, could benefit from such patterns.

The results show that our NLP approach is able to effectively learn writing styles revealed from syntax, semantics, latent topics, word and character *n*-grams to determine expertise levels. This study also demonstrated the adaptability and extensibility of the proposed NLP approach. That implies the approach potentially allows further enhancements via enrichment of existing linguistic annotations.

Even though our experiments and results are based on world's leading and largest crowdsourcing forums (Zooniverse and StackExchange), it is interesting to include other related forums such as Yahoo!Answers, Quora and so on.

About 20 different **expertise models** have been built on the selected linguistic annotations. The *baseline* models, which are **unigram latent topic words based**, have been iteratively *enhanced* in a *two-stage process*, by adding syntactic, semantic, word and character *n*-gram punctuation and character set annotations. That demonstrates the **extensibility** of the models to adapt other types of linguistic features.

Model validation and evaluation have been carried out on **8 crowdsourced corpora**: 2 of them are **in-domain test sets**, the remaining datasets are **related** and **out-domain**. The evaluation results on such quite heterogeneous collections of data, guarantee the validity of the learned models across domains. The best model yields nearly **0.7940 R-squared** and **0.0263** and **0.2200**, **RMSE**, range and mean normalized, respectively.

Standard techniques widely used in NLP are applied in this study. In addition to such techniques, in the future, we are also interested to further benefit from little or none exploited, but quite profound and deep text analysis methods [55]. That revealed many interesting truths hidden in various sources of text. For instance, Joshua in Ref. [55], suggest story structuring strategies.

The **expertise dimensions** identified and defined using syntactic structure analysis, in this study, could possibly be extended to embrace other important parameters, possibly via semantic analysis. Eventually, establish a standard **concrete framework**, which could serve many CSFs for the evaluation of users' competence. In addition, for the future, it would be interesting to advance the method for capturing **code snippets-related information**, to better evaluate the quality of programmers' contributions with in question-answering communities. Combining our growing effort of enhancing media searchability with studies that particularly focus on improving media understandability [19, 56] might also be interesting for the future.

Acknowledgements We would like to acknowledge the Zooniverse team at Oxford University and StackExchange, for providing the experimental data.

Funding Open access funding provided by Umea University.

Declarations

Conflict of interest The author states that there is no conflict of interest.

References

1. Tausczik Y, Pennebaker J. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol. 2010;29(1):24–54.
2. Bryan S, Robert H. Analyzing cockpit communications: the links between language, performance, error, and workload. Hum Perform Extreme Environ. 2000;5(1):63–8.
3. Baltadzhieva A, Chrupala G. Question quality in community question answering forums: a survey. SIGKDD Explor. 2015;17:8–13.
4. Chen Y, Wrenn JO, Xu H, Spickard A, Habermann R, Powers JS, et al. Automated assessment of medical students' clinical exposures according to AAMC geriatric competencies. In: AMIA annual symposium proceedings AMIA symposium; 2014. Vol. 2014, p. 375–384.
5. Woldemariam Y. Assessing users' reputation from syntactic and semantic information in community question answering. In: Proceedings of the 12th conference on language resources and evaluation (LREC 2020); 2020. p. 5385–5393.
6. Aichroth P, Weigel C, Kurz T, Stadler H, Drewes F, Björklund J, et al. MICO-media in context. In: Proceedings of 2015 IEEE international conference on multimedia and expo workshops (ICMEW); 2015. p. 1–4.
7. Le N, Bredin H, Sargent G, India M, Lopez-Otero P, Barras C, et al. Towards large scale multimedia indexing: a case study on person discovery in broadcast news. In: Proceedings of international workshop on content-based multimedia retrieval; 2017. p. 1–6.
8. Woldemariam Y. Sentiment analysis in a cross-media analysis framework. In: 2016 IEEE international conference on big data analysis (ICBDA); 2016. p. 1–5.
9. Woldemariam Y, Dahlgren A. Adapting language specific components of cross-media analysis frameworks to less-resourced languages: the case of Amharic. In: Proceedings of the 1st joint workshop on spoken language technologies for under-resourced languages (SLTU) and collaboration and computing for under-resourced languages (CCURL); 2020. p. 298–305.
10. Woldemariam Y. Transfer learning for less-resourced semitic languages speech recognition: the case of Amharic. In: Proceedings of the 1st joint workshop on spoken language technologies for under-resourced languages (SLTU) and collaboration and computing for under-resourced languages (CCURL); 2020. p. 61–69.
11. MacLeod L. Reputation on stack exchange: tag, you're it! In: Proceedings of the 28th international conference on advanced information networking and applications workshops; 2014. p. 670–674.
12. Movshovitz-Attias D, Movshovitz-Attias Y, Steenkiste P, Faloutsos C. Analysis of the reputation system and user contributions on a question answering website: StackOverflow. In: 2013 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2013); 2013. p. 886–893.
13. Cai Y, Chakravarthy S. Expertise ranking of users in QA community. In: DASFAA; 2013. p. 25–40.
14. Zhang J, Ackerman MS, Adamic L. Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on world wide web; 2007. p. 221–230.
15. Shah C, Pomerantz J. Evaluating and predicting answer quality in community QA. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval; 2010. p. 411–418.
16. Zhu Z, Bernhard D, Gurevych I. A multi-dimensional model for assessing the quality of answers in social QA sites. Technische Universitat Darmstad; 2009. Technical Report TUD-CS-2009-0158.
17. Gauthier J, Levy R, Tenenbaum JB. Word learning and the acquisition of syntactic–semantic overhypotheses. arXiv preprint arXiv:180504988. 2018.
18. Tausczik Y, Pennebaker J. Predicting the perceived quality of online mathematics contributions from users' reputations. In: Proceedings of the SIGCHI conference on human factors in computing systems; 2011. p. 1885–1888.
19. Chang YC, Ku CH, Chen CH. Social media analytics: extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. Int J Inf Manag. 2019;48:263–79.
20. Woldemariam Y, Björklund H, Bensch S. Predicting user competence from linguistic data. In: Proceedings of the 14th international conference on natural language processing (ICON-2017); 2017. p. 476–484.
21. Greg D, Adam P, Dan K. Syntactic transfer using a bilingual lexicon. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning; 2012. p. 1–11.
22. Wang D, Zheng TF. Transfer learning for speech and language processing. In: Asia-pacific signal and information processing association annual summit and conference (APSIPA); 2015. Vol. 2015, p. 1225–1237.
23. Ghoshal A, Swietojanski P, Renals S. Multilingual training of deep neural networks. IEEE Int Conf Acoust Speech Signal Process. 2013;2013:7319–23.
24. Huang JT, Li J, Yu D, Deng L, Gong Y. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. IEEE Int Conf Acoust Speech Signal Process. 2013;2013:7304–8.
25. Porter MF, et al. An algorithm for suffix stripping. Program. 1980;14(3):130–7.
26. Klein D, Manning CD. Accurate unlexicalized parsing. In: ACL; 2003. p. 423–443.
27. Zhu M, Zhang Y, Chen W, Zhang M, Zhu J. Fast and accurate shift-reduce constituent parsing. In: Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers); 2013. p. 434–443.
28. Ján S, Pavol N. Full text search engine as scalable k-nearest neighbor recommendation system. Berlin: Springer; 2010.
29. Wallach HM. Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning; 2006. p. 977–984.
30. Zhang J, Gong S. Action categorization by structural probabilistic latent semantic analysis. Comput Vis Image Underst. 2010;114(8):857–64.
31. Correa D, Sureka A. Chaff from the wheat: characterization and modeling of deleted questions on stack overflow. In: Proceedings of the 23rd international conference on world wide web; 2014. p. 631–642.
32. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: Advances in neural information processing systems; 2015. p. 649–657.
33. Mcnamee P, Mayfield J. Character n-gram tokenization for European language text retrieval. Inf Retr. 2004;7(1–2):73–97.
34. Freedman D. Statistical models: theory and practice. Cambridge: Cambridge University Press; 2005.
35. Yan X, Su X. Linear regression analysis: theory and computing. New Jersey: World Scientific; 2009.
36. Alexopoulos EC. Introduction to multivariate regression analysis. Hippokratia. 2010;14(Suppl 1):23–8.
37. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci Model Dev. 2014;7(3):1247–50.
38. Shcherbakov MV, Brebels A, Shcherbakova NL, Tyukov AP, Janovsky TA, Kamaev VA. A survey of forecast error measures. World Appl Sci J. 2013;24(24):171–6.

39. Baltadzhieva A, Chrupala G. Predicting the quality of questions on Stackoverflow. In: Proceedings of the international conference recent advances in natural language processing; 2015. p. 32–40.
40. Garten J, Boghrati R, Hoover J, Johnson KM, Dehghani M. Morality between the lines: detecting moral sentiment in text. In: Proceedings of IJCAI 2016 workshop on computational modeling of attitudes; 2016.
41. Schwarz N. Self-reports: how the questions shape the answers. Am Psychol. 1999;54(2):93–105.
42. Zhang L, Huang X, Liu T, Li A, Chen Z, Zhu T. Using linguistic features to estimate suicide probability of Chinese microblog users. In: International conference on human centered computing; 2014. p. 549–559.
43. Artzi Y, Pantel P, Gamon M. Predicting responses to microblog posts. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies; 2012. p. 602–606.
44. Liu X, Zhu T. Deep learning for constructing microblog behavior representation to identify social media user's personality. PeerJ Comput Sci. 2016;2:e81.
45. Merity S, Keskar NS, Socher R. Regularizing and optimizing LSTM language models. arXiv preprint arXiv:170802182. 2017.
46. Kneser R, Ney H. Improved backing-off for m-gram language modeling. In: 1995 International conference on acoustics, speech, and signal processing; 1995. Vol. 1, p. 181–184.
47. El-Din DM. Enhancement bag-of-words model for solving the challenges of sentiment analysis. Int J Adv Comput Sci Appl. 2016;7(1).
48. Voorhees EM. Natural language processing and information retrieval. In: International summer school on information extraction; 1999. p. 32–48.
49. Agarwal A, Biadsy F, Mckeown K. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In: Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009); 2009. p. 24–32.
50. Houvardas J, Stamatatos E. N-gram feature selection for authorship identification. In: International conference on artificial intelligence: methodology, systems, and applications; 2006. p. 77–86.
51. Marneffe MC, MacCartney B, Manning C. Generating typed dependency parses from phrase structure parses. In: Proc. 5th international conference on language resources and evaluation (LREC 2006); 2006. p. 449–454.
52. Nivre J, de Marneffe MC, Ginter F, Goldberg Y, Hajic J, Manning CD, et al. Universal dependencies v1: a multilingual treebank collection. In: LREC; 2016. p. 1659–1666.
53. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations; 2014. p. 55–60.
54. Crossley S, Liu R, McNamara D. Predicting math performance using natural language processing tools. In: Proceedings of the seventh international learning analytics; knowledge conference; 2017. p. 339–347.
55. Schimel J. Writing science: how to write papers that get cited and proposals that get funded. OUP USA; 2012.
56. Gupta MS, Kumar K. Progression on spectrum sensing for cognitive radio networks: a survey, classification, challenges and future research issues. J Netw Comput Appl. 2019;143:47–76.