# Characterizing discussions in the Spanish Wikipedia

Johnny Torres*, Alfonsina Ochoa*, Alberto Jimenez*, Sixto García*, Enrique Peláez*, Xavier Ochoa*

* Escuela Superior Politécnica del Litoral, ESPOL,
Facultad de Ingeniería Eléctrica y Computación,
Centro de Tecnologías de Información, CTI,
Campus Gustavo Galindo Km 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
{johnny.torres, alfonsina.ochoa, alberto.jimenez, sgarcia, epelaez, xavier}@cti.espol.edu.ec

***Abstract*—Wikipedia, as the largest online encyclopedia, is edited collaboratively by hundreds of users. The content in some articles can have dispute, giving rise to discussions which are registered in the related talk pages. In this paper, we propose an annotation schema for Spanish Wikipedia talk pages in order to determine the type of opinions expressed in them. We apply the annotation schema to a corpus that includes a collection of discussions about 148 topics drawn from 25 Spanish Wikipedia talk pages. We make the resulting dataset publicly available for download on github[1]. Furthermore, we train and evaluate supervised machine learning models to automatically identify the annotation labels. Linear Support Vector classifier (LinearSVC) performs better compared to other baseline models, and achieves an accuracy $F_1 = 0.71$ in our experiments.**

***Index Terms*—Wikipedia; Collaborative Writing; NLP.**

## I. Introduction

Nowadays, the paradigm of generating content in the web has shifted from individual to collaborative content production. Wikipedia is a free online encyclopedia, widely used for creating articles collaboratively, where many people around the world are constantly improving its content. Also, it allows to discuss any article's contents on corresponding talk pages.

The open nature of the Wikipedia encyclopedia has accelerated the growth of its contents to more than 45 million articles in 293 languages[2], until June 2017. It also has enabled researchers to study several aspects of collaborative writing of articles, such as: writing patterns, quality, interactions, and the role of its users (wikipedians). Although, these aspects have been studied extensively for languages like English or German, the Spanish Wikipedia has received less attention [1].

The discussion pages on Wikipedia, formally known as talk pages, serve as a mean for communication and coordination. Each article can have a talk page associated. On these pages, the wikipedians can participate in discussions about the content of the article. For instance, they can discuss the modifications to be made on the article, including sections to be deleted or rewritten [2]. Wikipedians are mostly editors, although some of them have roles as reviewers or coordinators. The latter roles are fundamental in organizing, explaining the community norms and policies, which facilitates information dissemination and improves the article quality [3].

In this paper, we aim to answer the following research question about the discussions pages in the Spanish Wikipedia: Can we automatically identify the type of opinions in discussions that describe the efforts to improve articles content quality?

As results of this work, our contributions are as follow: 1) an annotation taxonomy for Spanish Wikipedia discussions describing the efforts to enhance the quality of related articles, 2) a corpus for the Spanish Wikipedia consisting of 2097 annotated opinions extracted from 25 talk pages, 3) a machine learning pipeline that uses a supervised classifier to achieve an accuracy score $F1$ of $0.71$. We made our dataset and code used in this paper[1] available to the research community.

## II. Related work

With the rise of the social web, the amount of studies analyzing user generated discussion significantly increased. In addition to exploring emails [4], web forums [5] and chats [6], Wikipedia talk pages have attracted the attention of researchers. These pages play an essential role as the place for discussion, collaboration and communication.

There is a wide range of computational work for classifying discussions and identifying social roles on Wikipedia. At the level of modeling the language and structure of the interaction exhibited by the participants on Wikipedia talk pages, researchers have endeavored to annotate social acts [7], to identify dialog acts [2], to identify disagreement and agreement expressions [8, 9] in users interactions. Also, it has been studied the roles of users that facilitate the coordination, moderation, and others tasks to improve the quality of articles [3, 10].

Social media platforms have enabled people from anywhere to express their points of view and discuss issues of interest. Laniado et al. offered an analysis of talk pages associated to articles and to users; they examined the Wikipedia discussion networks in order to capture patterns of interaction and created tree structures of the discussion. The interpretation of the graphs revealed patterns that are unique to Wikipedia discussions and suggested some metrics that might be used to characterize different types of talk pages.

Other studies have introduced text classification taxonomies. For example, Bender et al. proposed a taxonomy considering

[1] https://github.com/espol-cti/cwdiscussions

[2] https://stats.wikimedia.org/EN/TablesWikipediaZZ.htm

978-1-5386-3894-1/17/$31.00 ©2017 IEEE

the two characteristic aspects of interaction on Wikipedia: authority claims and alignment moves. They analyzed the interactions between participant status and social acts; how the participants established their credibility in the discussions and how they expressed disagreement and agreement towards other participants or topics. From a different perspective, Yang et al. [10] examined Wikipedia discussion networks in order to identify the editing interactions, and introduced a fine-grained taxonomy of edit types to train machine learning models to automatically identify editing actions.

Our study is closely related to the latter studies. To the best of our knowledge, there is no work yet that establishes the opinion types for discussions on Spanish Wikipedia.

## III. Dataset

### A. Source Data

Open source projects, like Wikipedia, provide valuable data for studies as the proposed in this paper. Besides collecting writing activities, Wikipedia captures the interactions of wikipedians in the form of discussions on discussion pages. Although, some wikipedians might interact using other communication channels, such as IRC or mailing lists, most interactions occur inside talk pages.

Each edition by wikipedians is stored in the Wikipedia database as a new page revision or version. Edit actions are not performed in real time, i.e. it does not offer real time collaboration like Google Docs or Word. Editing pages in Wikipedia is similar to the behavior in version control systems used in software development (e.g. Git, Subversion) where the users make check-out, perform changes in the document, and finally make check-in of the set of changes as a whole.

Additionally, for all Wikipedia pages (including talk pages), the system captures additional metadata attributes in editing actions, such as: identity of wikipedians, date and time, and special markups. The identity of wikipedians can be the username for registered users or IP address in the case of unregistered users. Special markups refer to formatting tags in the page content, and hyperlinks to other pages or external sources. These metadata can be used for several more detailed analyses at editor or page level.

The Wikimedia Foundation periodically releases dumps of the Wikipedia database in XML format. These dumps are separated by language and contains the current version and entire edit history for each page. For purposes of this study, we collected the dump of Spanish Wikipedia published in June 2017, which is freely available[3]. The table I shows the number of pages found in Spanish Wikipedia, with 51% corresponding to *articles pages* namespace[4]. For articles, 31% of them has a *talk page* associated.

As shown in table I, not all articles have a discussion page associated. Recent articles account for a percentage of the missing talk pages, but the vast majority is due to articles with low quality or relevance. These articles do not attract enough wikipedians' attention to discuss or coordinate actions for improving them.

TABLE I: Spanish Wikipedia Corpus

| Pages | Articles | Talks |
|---|---|---|
| 5,918,915 | 1,297,351<br>22% | 404,367<br>31% |

Our goal is to analyze the talk pages associated to Wikipedia articles. Therefore, we defined a list of articles to extract from the XML dump for further processing. Articles in Wikipedia are associated to one or more categories, and for the Spanish version, we found that 99.96% of the articles belong to at least one category. The list of selected pages includes articles in the category of biographies (C2: man, C3: living persons), which are categories that have a high number of pages and discussions, as shown in figure 1. The correlation between the number of articles and talk pages in the top ten categories is 98.18%.

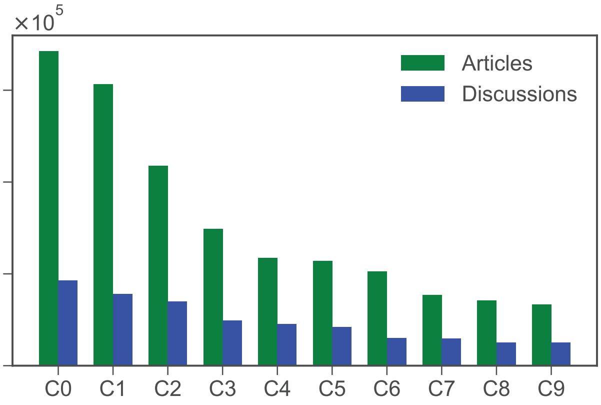


Fig. 1: Top categories by number of articles.

From the category of biographies, we have chosen talk pages of political leaders in the American continent, that are in office or has been presidents recently. The reason of this selection criteria is that those articles have a lot of discussions or activities in their corresponding talk pages. Our dataset contains 25 talk pages, edited 10942 times, by 2922 wikipedians, from May 10, 2004 to May 30, 2017.

### B. Segmentation

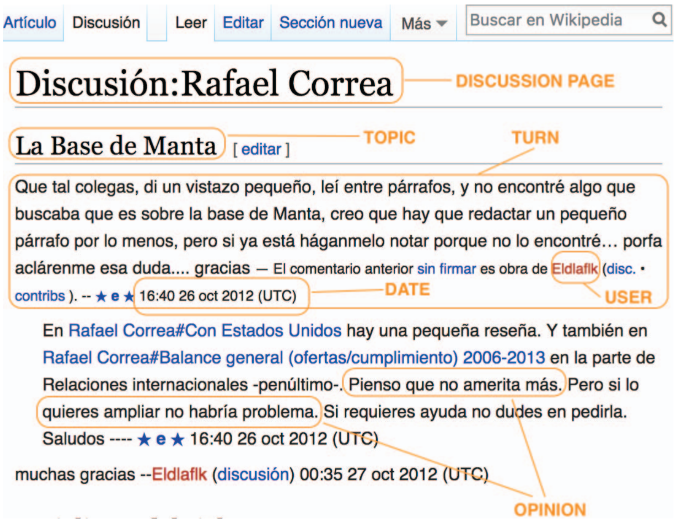
In order to read the dumps and segment the content of talk pages, we used the methodology proposed in [12]. Thus, we determine the precise authorship of each word in the selected talk pages. Previous studies have segmented Wikipedia talk pages down to the level of *turns* [7, 2]. At this level, turns represent the consecutive body of text written by a wikipedian in a single edit, until another wikipedian replies in the same discussion thread. Usually, turns are associated to a paragraph, and the replies are often indented. Moving up in the hierarchy of segmentation, each turn belongs to a specific topic of

[3] https://dumps.wikimedia.org/eswiki/latest/

[4] https://en.wikipedia.org/wiki/Wikipedia:Namespace

discussion. Finally, each topic in the talk pages is related to a section or aspect of its corresponding main article.

Our approach is to segment each article down to the level of *opinions*, i.e. sentences or phrases, rather than *turns*. The intuition behind this criterion is that: a wikipedian can express several ideas in each turn. Figure 2 shows an example of the discussion page about the article of the former President Rafael Correa in our corpus. One topic of discussion is ”La base de Manta” and it has three turns, the second turn contains more than one opinion.



Discusión:Rafael Correa

La Base de Manta [ editar ]

Que tal colegas, di un vistazo pequeño, leí entre párrafos, y no encontré algo que buscaba que es sobre la base de Manta, creo que hay que redactar un pequeño párrafo por lo menos, pero si ya está háganmelo notar porque no lo encontré… porfa aclárenme esa duda.... gracias — El comentario anterior sin firmar es obra de Eldlaflk (disc. • contribs ). -- ★ e ★ 16:40 26 oct 2012 (UTC)

En Rafael Correa#Con Estados Unidos hay una pequeña reseña. Y también en Rafael Correa#Balance general (ofertas/cumplimiento) 2006-2013 en la parte de Relaciones internacionales -penúltimo-. Pienso que no amerita más. Pero si lo quieres ampliar no habría problema. Si requieres ayuda no dudes en pedirla. Saludos ---- ★ e ★ 16:40 26 oct 2012 (UTC)

muchas gracias --Eldlaflk (discusión) 00:35 27 oct 2012 (UTC)

Fig. 2: Spanish Wikipedia talk page of Rafael Correa.

We have considered special cases of segmentation to minimize errors during the process. A special case is when a wikipedian can create multiple turns in a single edit, or a turn is created in multiple edit actions. These cases correspond to 5% of turns segmented. We found another special case when multiple wikipedians modify the same opinion, this occurs in 3% of the opinions. This last case is an exception in talk pages, compared to articles, where often several wikipedians can modify to the same sentence. For the segmentation process, we selected only the first 1000 revisions of each talk page resulting in 7587 opinions that belong to 2438 turns, created by 739 wikipedians.

The analysis of the automatic segmentation detected errors in 6% of the opinions, which were fixed manually. Usually, the wikipedian signatures can be found at the end of each turn. Although some signatures are preceded by –, others do not use any specific separator. Also, we found three opinions written in a different language than Spanish, and we flag those as type OLAN in our corpus.

### *C. Annotation*

To annotate the type of opinions expressed by wikipedians, we defined a taxonomy that captures the action-oriented message to improve the associated main article. Based on the taxonomies proposed in [7, 2], we create a higher level of abstraction to group those annotation labels in three categories, as described next:

- **Argumentative**: denotes articles containing criticism or authority claims expressed regarding the content of the main article. Arguments from authority are grounded on work of many philosophers [13, 14], and recent work on behavior of users on the Web [15, 16]. Many aspects (e.g. context, medium, genre) have been identified that may affect the way individuals react or take authority claims. Argumentation also include criticism to the content contributed by others wikipedians, as defined in [17, 2]. The following subcategories were considered for argumentative category: a) criticism to incomplete or missing details (CM), lack of accuracy or correctness (CW), unsuitable or unnecessary content (CU), structural problems (CS), deficiencies in language or style (CL), objectivity or bias issues (COBJ), or other kind of article criticism (CO). b) experiential based authority (ACE), credential based authority (ACC), institutional based authority (ACI), forum based authority (ACF), external based authority (ACEX), authority based on social expectations (ACSE).
- **Performative**: describes actions to take in order to implement edit activities based on argumentative opinions or by decisions of experienced wikipedians [2]. This category includes two kinds of subcategories: those requesting explicit actions to modify the content of articles, and implicit suggestions for information interchange: a) explicit suggestions or requests (PSR), reference or pointers (PREF), commitment to an action in the future (PFC), report of a performed action (PPC). b) information providing (IP), information seeking (IS), information correcting (IC).
- **Interpersonal**: denotes attitudes between wikipedians. These attitudes can be: a) positive attitude or acceptance towards others, such as: greeting, thanking (ATT+). b) partial acceptance or partial rejection to other wikipedians (ATTP). c) negative attitude or rejection towards others, such as: mocking, bullying, or insulting (ATT-).

After the segmentation process in our corpus, we selected the first 100 opinions of each talk page for annotation. The annotation process was conducted by two external users, following specific guidelines and the proposed taxonomy. Then, we merged the annotation into a single file that can be used in our machine learning experiments. A third user was responsible of resolving disagreements.

In our methodology, the segmentation is applied at the opinion level because we found that turns can have multiple opinions in our corpus. Based on the results of our annotation process, we calculated the distribution of labels by turn. To establish the author which an opinion belongs to, we aggregated the opinions by author and considering a time span of maximum 5 minutes between edits in the same talk page. We found that more than 47% of the *turns* have more than one label associated as shown in the table II. This is in contrast to previous works [7, 2], where each turn has associated only one label.

Because of the complexity of the annotation process for

TABLE II: Turns labels distribution

| **# of labels** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **%** | 53.31% | 26.93% | 11.60% | 5.39% | 2.21% | 0.28% | 0.14% | 0.14% |

unstructured text and data found in Wikipedia talk pages, it is important to examine the quality of the annotation process. Various quantitative metrics have been proposed to measure the underlying consistency of annotation process [18]. Table III shows the two metrics of quality that we calculate for each opinion label. The *agreement* metric $p_o$ defines the percentage of opinions in which both annotators agreed. The more robust metric kappa $\kappa$ defines an inter-annotator agreement coefficient that considers the probability of agreement by chance [19], which enforce a higher quality annotation. It is defined as:

$$\kappa = (p_o - p_e)/(1 - p_e) \tag{1}$$

where $p_o$ represents the empirical probability of agreement, i.e. the observed agreement ratio, and $p_e$ defines the expected agreement for randomly assigned labels by both annotators. For $p_e$, we estimated using the weight of each class label.

TABLE III: Agreement score

| subtype | N | po | pe | kappa |
|---|---|---|---|---|
| ACE | 3 | 0.67 | 0 | 0.67 |
| ATT | 148 | 0.99 | 0.16 | 0.99 |
| CL | 12 | 0.58 | 0.33 | 0.38 |
| CM | 58 | 0.81 | 0.19 | 0.77 |
| CO | 617 | 0.92 | 0.2 | 0.89 |
| COBJ | 29 | 0.79 | 0.14 | 0.76 |
| CS | 20 | 0.85 | 0.4 | 0.75 |
| CU | 36 | 0.67 | $8.33 \cdot 10^{-2}$ | 0.64 |
| CW | 36 | 0.83 | 0.22 | 0.79 |
| IC | 14 | 0.79 | 0.36 | 0.67 |
| IP | 126 | 0.98 | 0.17 | 0.98 |
| IS | 22 | 0.64 | 0.32 | 0.47 |
| PFC | 16 | 0.88 | 0.19 | 0.85 |
| PPC | 75 | 0.88 | 0.15 | 0.86 |
| PREF | 124 | 0.89 | 0.2 | 0.86 |
| PSR | 243 | 0.99 | 0.2 | 0.98 |

The reliability of annotation has been addressed for other datasets in the scope of related work. Kim et al. examined 1135 post of students' discussions in online forums with five labels. They reported a high Kappa score, between 0.72 and 0.94, probably due to their coarse-grained label taxonomy. Ferschke et al. reported a kappa score between 0.13 and 0.66, for 365 discussions in Wikipedia talk pages. Ferschke et al. reported their kappa score ranging from 0.18 to 0.92. The last two works are closely related to our study, and our results present a similar dispersion in the agreement scores due to the fine-grained taxonomy.

## IV. Identifying Opinions

In this study, we aim to identify the type of opinions in Wikipedia talk pages. Prior to describe the classification models, we first analyze several aspects of the metadata and text of opinions in our corpus.

### A. Feature extraction

**Opinions**. We analyze the patterns of users contributing in the discussions. We found that 92% of the users contributed to only one specific talk page. In our corpus, few users participated in the discussions of two or more different articles, but it does not affect the distribution of users across discussion. We also found that the number of opinions contributed by wikipedians follows a power law distribution. That means most users contributed a small number of opinions and it decays to a few very active users. The number of opinions by pages has a long right tail distribution, due to some articles having few discussions. Nonetheless, the majority is close to the sampling size defined in the annotation process.

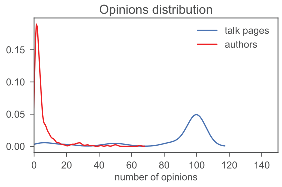


Fig. 3: Opinions distribution.

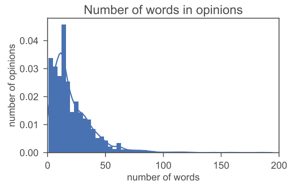


Fig. 4: Words distribution.

**Temporality**. The timeline of the opinions contributed by wikipedians can be observed at the bottom of figure 5. We analyze the spikes present in our corpus. At the top of figure 5, we extract the word cloud for the most prominent spike from Jan 1, 2015 to Jan 1, 2017. To obtain the word cloud, we remove outliers in the words distribution. We found that the topics are related to specific events in the context of the political activity of each leader, mainly during elections. We take in consideration these temporal outliers to prevent overfitting in our classification model.

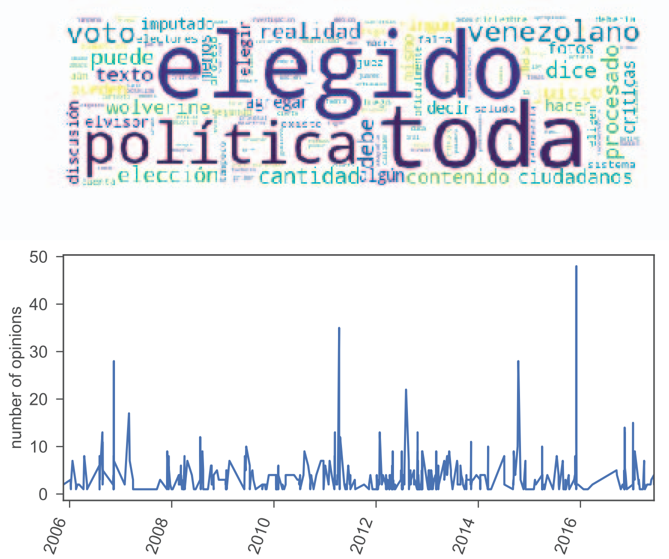


Fig. 5: Temporal distribution of opinions contributed by wikipedians.

**Text**. In order to obtain a good performance in our classification task, we analyze the text of opinions in talk pages. The distribution number of characters and words by opinions

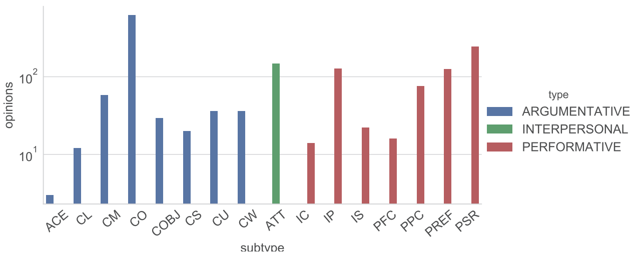


Fig. 6: Distribution of annotated opinions.

(figure 4) shows that most of the opinions are short, following a power law distribution. Prior to feed the opinion text to the learning algorithms, we performed a preprocessing task using methods in [21] that includes: 1) Removal stops words, Spanish accents, URLs, numbers, dates. 2) Removal of infrequent words, with a minimum frequency of three. 3) Stemming of each word, normalize each word to its root.

### B. *Classification*

We define our problem as a multi-class classification problem, where each opinion type is a mutually exclusive class [22]. We train and evaluate two supervised learning models using the toolkit developed by Pedregosa et al., 2011.

We use the words unigrams from the opinions as content features. These unigrams are vectorized using a logarithmic inverse TF-IDF [24]. We split our dataset, assigning $65\%$ of the opinions for training and $35\%$ for testing purposes. To account for the imbalance in the labels, we use stratified splitting strategy.

**Labels**. As a result of the annotation process, figure 6 shows the labels distribution. There is an unbalance, with some opinion types with less than 50 instances. In this study, we use the higher level of our taxonomy for the classification task.

## V. Results

For our experiments, a uniform random classifier was used to establish the baseline. We evaluated two other algorithms commonly used in text classification tasks:

- Multinomial Naive Bayes (MNB) which is suitable for sparse data classification[25].
- Linear SVC (LSVC) supports both dense and sparse data, and it uses a one-vs-the-rest scheme[26].

We applied a randomized grid search for hyper-parameters tuning because of the performance improvements over an exhaustive grid search [27]. Then, we applied a cross validation over the training set to establish the precision score with a confidence level of $95\%$ as shown in table IV . The results reflect the complexity of this task, as the baseline is very low.

Table IV shows the performance of the classification models. The column *Label* indicates the opinion types present in our data filtered and preprocessed that feed the supervised learning algorithms. The column *Support* indicates the number of instances of each class present in the test set. The following columns correspond to the precision, recall, and F1 scores of the learning algorithms [28]. The scores were calculated at micro level, i.e. for each class or opinion type, and at macro level. Linear SVC shows better overall scores compared to other baseline algorithms: slightly better precision that MNB, and more robust recall and F1. Our classification approach achieves an accuracy $F_1 = 0.71$. The ability to automatically classify discussion pages will help to investigate the relations between article discussions and article edits, which is an important step towards understanding the processes of collaboration on Wikipedia.

## VI. Conclusions

The discussions on Wikipedia can provide useful insights about the editing process for collaborative writing of articles, and its relation to the quality improvement. The classification of opinions types in online discussion, particularly on Wikipedia talk pages, is a complex task that often requires understanding of non-explicit content like meaning and context. In this study, we propose an annotation schema and a baseline classification pipeline for identifying type of opinions in discussions on Spanish Wikipedia. The proposed model is able to identify types of opinions expressed by wikipedians in three high-level categories.

Future research directions include improving our corpus at low-level categories. Generating large number of labels can to be a costly and challenging task, therefore unsupervised or semi-supervised learning is an important research direction. It would be interesting to analyze other categories on Wikipedia, and find out if the behavior of wikipedians change. Another aspect of interest could be to uncover the roles of wikipedians and how they react in different context scenarios, such as, location or gender. In the area of the machine learning, it would be important to evaluate the domain adaptation for additional categories using external data sources. Also, a deeper understanding of meaning, sentiment, and stance is needed in order to fully understand the role and motivation behind wikipedians to improve the articles quality.

## References

[1] T. K. Park, "The visibility of wikipedia in scholarly publications," *First Monday*, vol. 16, no. 8, 2011.

[2] O. Ferschke, I. Gurevych, and Y. Chebotar, "Behind the article: Recognizing dialog acts in wikipedia talk pages," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 777–786.

[3] X. Qin, P. Cunningham, and M. Salter-Townshend, "The influence of network structures of wikipedia discussion pages on the efficiency of wikiprojects," *Social Networks*, vol. 43, pp. 1–15, 2015.

[4] W. W. Cohen, V. R. Carvalho, and T. M. Mitchell, "Learning to classify email into" speech acts"." in *EMNLP*, vol. 4, 2004, pp. 309–316.

[5] S. N. Kim, L. Wang, and T. Baldwin, "Tagging and linking web forum posts," in *Proceedings of the Four-*

TABLE IV: Classification models comparison

| Label | Support | BL-P | BL-R | BL-F1 | LSVC-P | LSVC-R | LSVC-F1 | MNB-P | MNB-R | MNB-F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| ARGUMENTATIVE | 284 | 0.48 | 0.33 | 0.39 | 0.71 | 0.73 | 0.72 | 0.63 | 0.85 | 0.72 |
| INTERPERSONAL | 52 | 0.09 | 0.33 | 0.14 | 0.86 | 0.71 | 0.78 | 1.00 | 0.65 | 0.79 |
| PERFORMATIVE | 217 | 0.34 | 0.27 | 0.30 | 0.64 | 0.65 | 0.64 | 0.65 | 0.41 | 0.51 |
| Macro | 553 | 0.30 | 0.31 | 0.28 | 0.74 | 0.70 | 0.71 | 0.76 | 0.64 | 0.67 |
| CI (95%) | | 0.05 | | | 0.06 | | | 0.02 | | |

*teenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2010, pp. 192–202.

[6] T. Carpenter and E. Fujioka, "The role and identification of dialog acts in online chat," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[7] E. M. Bender, J. T. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, and M. Ostendorf, "Annotating social acts: Authority claims and alignment moves in wikipedia talk pages," in *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011, pp. 48–57.

[8] L.-M. Ho-Dac, V. Laippala, C. Poudat, and L. Tanguy, "French wikipedia talk pages: Profiling and conflict detection," in *4th Conference on CMC and Social Media Corpora for the Humanities*, 2016.

[9] L. Wang and C. Cardie, "Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon," *arXiv preprint arXiv:1606.05706*, 2016.

[10] D. Yang, A. Halfaker, R. E. Kraut, and E. H. Hovy, "Edit categories and editor role identification in wikipedia." in *LREC*, 2016.

[11] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner, "When the wikipedians talk: Network and tree structure of wikipedia discussion pages." in *ICWSM*, 2011, pp. 177–184.

[12] F. Flöck and M. Acosta, "Wikiwho: Precise and efficient attribution of authorship of revisioned content," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 843–854.

[13] J. Derrida, "Force of law: The'mystical foundation of authority'. in ed. drucilla cornell, michael rosenfield and david g. carlson," 1992.

[14] Y. Liu, "Authority, presumption, and invention," *Philosophy & rhetoric*, vol. 30, no. 4, pp. 413–427, 1997.

[15] B. Amento, L. Terveen, and W. Hill, "Does "authority" mean quality? predicting expert quality ratings of web documents," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000, pp. 296–303.

[16] J. L. Jensen, "Public spheres on the internet: Anarchic or government-sponsored–a comparison," *Scandinavian political studies*, vol. 26, no. 4, pp. 349–374, 2003.

[17] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Information quality work organization in wikipedia," *Journal of the Association for Information Science and Technology*, vol. 59, no. 6, pp. 983–1001, 2008.

[18] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[19] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[20] J. Kim, J. Li, and T. Kim, "Towards identifying unresolved discussions in student online forums," in *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2010, pp. 84–91.

[21] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[22] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[25] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[26] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[27] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[28] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.