Controversy in Context

Abstract

With the growing interest in social applications of Natural Language Processing and Computational Argumentation, a natural question is how controversial a given concept is. Prior works relied on Wikipedia's metadata and on content analysis of the articles pertaining to a concept in question. Here we show that the immediate textual context of a concept is strongly indicative of this property, and, using simple and language-independent machine-learning tools, we leverage this observation to achieve state-of-the-art results in controversiality prediction. In addition, we analyze and make available a new dataset of concepts labeled for controversiality. It is significantly larger than existing datasets, and grades concepts on a 0-10 scale, rather than treating controversiality as a binary label.

Introduction

Indicating that a web page is controversial, or disputed - for example, in a search result - facilitates an educated consumption of the information therein, suggesting the content may not represent the “full picture”. Here, we consider the problem of estimating the level of controversiality associated with a given Wikipedia concept (title). We demonstrate that the textual contexts in which the concept is referenced can be leveraged to facilitate this.

The definition of which concepts are controversial is controversial by itself; an accurate definition of this elusive notion attracted the attention of researchers from various fields, see for example some recent attempts in BIBREF0, BIBREF1, BIBREF2.

Most people would agree, for example, that Global warming is a controversial concept, whereas Summer

is not. However, the concept Pollution may be seen as neutral by some, yet controversial by others, who associate it with environmental debates. In other words, different people may have different opinions, potentially driven by different contexts salient in their mind. Yet, as reported in the sequel, an appreciable level of agreement can be reached, even without explicit context.

Focusing here on Wikipedia concepts, we adopt as an initial “ground truth” the titles listed on the Wikipedia list of controversial issues, which is curated based on so-called “edit wars”. We then manually annotate a set of Wikipedia titles which are locked for editing, and evaluate our system on this much larger and more challenging dataset.

To estimate the level of controversy associated with a Wikipedia concept, we propose to simply examine the words in the sentences in which the concept is referenced. Because a concept can often be found in multiple contexts, the estimation can be seen as reflecting the “general opinion” about it in the corpus. This contrasts previous works, which consider this a binary problem, and employ a complex combination of features extracted from Wikipedia's article contents and inter-references, and more extensively – from the rich edit history thereof.

Related work

Analysis of controversy in Wikipedia, online news and social media has attracted considerable attention in recent years. Exploiting the collaborative structure of Wikipedia, estimators of the level of controversy in a Wikipedia article were developed based on the edit-history of the article BIBREF0, BIBREF3. Along these lines, BIBREF4 detect controversy based on mutual reverts, bi-polarity in the collaboration network, and mutually-reinforced scores for editors and articles. Similarly, BIBREF1 classify whether a Wikipedia page is controversial through the combined evaluation of the topically neighboring set of pages.

Content analysis of controversial Wikipedia articles has been used to evaluate the level of controversy of other documents (e.g., web pages) by mapping them to related Wikipedia articles BIBREF5. BIBREF6 further build a language model, which enhances predictions made by existing classifiers, by inferring word probabilities from Wikipedia articles prominent in Wikipedia controversy features (mainly signals in edit history as discussed above) and from articles retrieved by manually selected query terms, believed to indicate controversy.

BIBREF7 detect controversy in news items by inspecting terms with excessive frequency in contexts containing sentiment words, and BIBREF8 study controversy in user comments of news articles using lexicons. Finally, BIBREF9 suggest that controversy is not a universal but rather a community-related concept, and, therefore, should be studied in context.

Here we measure a concept's controversiality from the explicit sentence-level context in which it is mentioned. In this, our approach is reminiscent of BIBREF10, who suggest a similar approach to detect abstract concepts.

Estimating a concept's controversiality level ::: Datasets

We consider three datasets, two of which are a contribution of this work.

Dataset I consists of 480 concepts previously analyzed in BIBREF1, BIBREF4. 240 are positive examples, titles from the Wikipedia list of controversial issues, and 240 are negative examples chosen at random and exclusive of the positives. Over this dataset, we compare the methodology suggested here to those reported by BIBREF1, BIBREF4. As the latter report overall accuracy of their binary prediction, we convert our controversiality estimates to a binary classification by classifying the higher-scored half as controversial, and the lower half as non-controversial.

Dataset II is based on a more recent version of the Wikipedia list of controversial issues (May 2017). As positive examples we take, from this list, all concepts which appear more than 50 times in Wikipedia. This leaves 608 controversial Wikipedia concepts. For negative examples, we follow BIBREF1, BIBREF4 and select a like number of concepts at random. Here too, since each concept only has a binary label, we convert our estimation into a binary classification, and report accuracy.

Dataset III is extracted from 3561 concepts whose Wikipedia pages are under edit protection, assuming that many of them are likely to be controversial. They were then crowd-annotated, with 10 or more annotators per concept. The annotation instructions were: “Given a topic and its description on Wikipedia, mark if this is a topic that people are likely to argue about.”. Average pairwise kappa agreement on this task was 0.532. Annotations were normalized to controversiality scores on an integer scale of 0 - 10. We used this dataset for testing the models trained on Dataset I.

In all datasets, to obtain the sentence-level context of the concepts (positive and negative), we randomly select two equal-sized sets of Wikipedia sentences, that explicitly reference these concepts – i.e., that contain a hyperlink to the article titled by the concept. Importantly, in each sentence we mask the words that reference the concept – i.e., the surface form of the hyperlink leading to the concept – by a fixed, singular token, thus focusing solely on the context within which the concepts are mentioned.

Estimating a concept's controversiality level ::: Controversiality Estimators

We employ three estimation schemes based on the textual contexts of concepts. The first relies on the context via pre-trained word embeddings of the concepts, which, in turn, are derived from the concepts' distributional properties in large samples of free texts. The other two schemes directly access the sentence-level contexts of the concepts.

Nearest neighbors (NN) Estimator: We used the pre-trained GloVe embeddings BIBREF11 of concepts to implement a nearest-neighbor estimator as follows. Given a concept c, we extract all labeled concepts within a given radius r (cosine similarity 0.3). In one variant, c's controversiality score is taken to be the fraction of controversial concepts among them. In another variant, labeled concepts are weighted by their cosine similarity to c.

Naive Bayes (NB) Estimator: A Naive Bayes model was learned, with a bag-of-words feature set, using the word counts in the sentences of our training data – the contexts of the controversial and non-controversial concepts. The controversiality score of a concept c for its occurrence in a sentence s, is taken as the posterior probability (according to the NB model) of s to contain a controversial concept, given the words of s excluding c, and taking a prior of 0.5 for controversiality (as is the case in the datasets). The controversiality score of c is then defined as the average score over all sentences referencing c.

Recurrent neural network (RNN): A bidirectional RNN using the architecture suggested in BIBREF10 was similarly trained. The network receives as input a concept and a referring sentence, and outputs a score. The controversiality score of a concept is defined, as above, to be the average of these scores.

Estimating a concept's controversiality level ::: Validation ::: Random @!START@k@!END@-fold

We first examined the estimators in k-fold cross-validation scheme on the datasets I and II with $k=10$: the set of positive (controversial) concepts was split into 10 equal size sets, and the corresponding sentences were split accordingly. Each set was matched with similarly sized sets of negative (non-controversial) concepts and corresponding sentences. For each fold, a model was generated from the training sentences and used to score the test concepts. Scores were converted into a binary classification, as described in SECREF3, and accuracy was computed accordingly. Finally, the

accuracy over the k folds was averaged.

Estimating a concept's controversiality level ::: Validation ::: Leave one category out

In a preliminary task, we looked for words which may designate sentences associated with controversial concepts. To this end, we ranked the words appearing in positive sentences according to their information gain for this task. The top of the list comprises the following: that, sexual, people, movement, religious, issues, rights.

The Wikipedia list of controversial issues specifies categories for the listed concepts, like Politics and economics, Religion, History, and Sexuality (some concepts are associated with two or more categories). While some top-ranked words - that, people, issues - do seem to directly indicate controversiality BIBREF12, BIBREF13, others seem to have more to do with the category they belong to. Although these categories may indeed indicate controversiality, we consider this as an indirect or implicit indication, since it is more related to the controversial theme than to controversiality per-se.

To control for this effect, we performed a second experiment where we set the concepts from one category as the test set, and used the others for training (concepts associated with the excluded category are left out, regardless of whether they are also associated with one of the training categories). We did this for 5 categories: History, Politics and economics, Religion, Science, and Sexuality. This way, thematic relatedness observed in the training set should have little or no effect on correctly estimating the level of controversy associated of concepts in the test set, and may even “mislead” the estimator. We note that previous work on controversiality does not seem to address this issue, probably because the meta-data used is less sensitive to it.

Results

Table TABREF14 compares the accuracy reported on Dataset I for the methods suggested in BIBREF1, BIBREF4 with the accuracy obtained by our methods, as well as the latter on Dataset II, using 10-fold cross-validation in all cases. Table TABREF14 reports accuracy results of the more stringent analysis described in section SECREF13.

BIBREF4 review several controversy classifiers. The most accurate one, the Structure classifier, builds, among others, collaboration networks by considering high-level behavior of editors both in their individual forms, and their pairwise interactions. A collaboration profile containing these individual and pairwise features is built for each two interacting editors and is classified based on the agreement or disagreement relation between them. This intensive computation renders that classifier impractical. Table TABREF14 therefore also includes the most accurate classifier BIBREF4 consider practical.

As seen in Table TABREF14, for the usual 10-fold analysis the simple classifiers suggested here are on par with the best and more complex classifier reported in BIBREF4. Moreover, in the leave-one-category-out setting (Table TABREF14), accuracy indeed drops, but only marginally. We also observe the superiority of classifiers that directly access the context (NB and RNN) over classifiers that access it via word embedding (NN).

Table TABREF14 presents results obtained when models trained on Dataset I are applied to Dataset III. For this experiment we also included a BERT network BIBREF14 fine tuned on Dataset I. The Pearson correlation between the scores obtained via manual annotation and the scores generated by our automatic estimators suggests a rather strong linear relationship between the two. Accuracy was computed as for previous datasets, by taking here as positive examples the concepts receiving 6 or more positive votes, and as negative a random sample of 670 concepts out of the 1182 concepts receiving no positive vote.

Conclusions

We demonstrated that the sentence–level context in which a concept appears is indicative of its controversiality. This follows BIBREF10, who show this for concept abstractness and suggest to explore further properties identifiable in this way. Importantly, we observed that this method may pick up signals which are not directly related to the property of interest. For example, since many controversial concepts have to do with religion, part of what this method may learn is thematic relatedness to religion. However, when controlling for this effect, the drop in accuracy is fairly small.

The major advantages of our estimation scheme are its simplicity and reliance on abundantly accessible features. At the same time, its accuracy is similar to state-of-the-art classifiers, which depend on complex meta-data, and rely on sophisticated - in some cases impractical - algorithmic techniques. Because the features herein are so simple, our estimators are convertible to any corpus, in any language, even of moderate size.

Recently, IBM introduced Project Debater BIBREF15, an AI system that debates humans on controversial topics. Training and evaluating such a system undoubtedly requires an extensive supply of such topics, which can be enabled by the automatic extraction methods suggested here as well as the new datasets.

Acknowledgment

We are grateful to Shiri Dori-Hacohen and Hoda Sepehri Rad for sharing their data with us and giving us permission to use it.