Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 11/1 (2018), 34-41
DOI: http://dx.doi.org/10.21609/jiki.v11i1.478

DETECTING CONTROVERSIAL ARTICLES ON CITIZEN JOURNALISM

Alfan Farizki Wicaksono, Sharon Raissa Herdiyana, and Mirna Adriani

Information Retrieval Lab., Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Email: alfan@cs.ui.ac.id

Abstract

Someone's understanding and stance on a particular controversial topic can be influenced by daily news or articles he consume everyday. Unfortunately, readers usually do not realize that they are reading controversial articles. In this paper, we address the problem of automatically detecting controversial article from citizen journalism media. To solve the problem, we employ a supervised machine learning approach with several hand-crafted features that exploits linguistic information, meta-data of an article, structural information in the commentary section, and sentiment expressed inside the body of an article. The experimental results shows that our proposed method manages to perform the addressed task effectively. The best performance so far is achieved when we use all proposed feature with Logistic Regression as our model (82.89% in terms of accuracy). Moreover, we found that information from commentary section (structural features) contributes most to the classification task.

Keywords: *controversy detection, text classification, supervised learning*

Abstrak

Pendirian dan pemahaman seseorang terhadap suatu topik kontroversial dipengaruhi oleh sumber berita yang dikonsumsinya. Namun, pembaca seringkali tidak menyadari bahwa ia sedang membaca sebuah artikel yang kontroversial. Padahal, dengan mengetahui bahwa sebuah artikel bersifat kontroversial, pembaca dapat lebih kritis dalam menerima informasi yang disampaikan di artikel tersebut. Penelitian ini bertujuan untuk mengembangkan sebuah model yang dapat secara otomatis mengklasifikasikan sebuah artikel jurnalisme warga berbahasa Indonesia sebagai kontroversial atau non-kontroversial. Digunakan metode berbasis *supervised learning* dengan dua model klasifikasi, yaitu *Logistic Regression* dan *Support Vector Machine*. Model dibangun dengan menggunakan empat kategori fitur, yaitu fitur metadata yang terdapat pada artikel, fitu struktural yang ada pada bagian komentar artikel, fitur linguistik, dan fitur yang mengeksploitasi informasi sentimen pada artikel. Hasil eksperimen menunjukkan bahwa model yang diusulkan oleh penelitian ini berhasil melakukan pendeteksian kontroversi dengan cukup efektif. Didapatkan akurasi terbaik sebesar 82,89% dengan menggunakan kombinasi semua fitur dan model *Logistic Regression*. Hasil eksperimen juga menunjukkan bahwa fitur struktural adalah fitur yang paling kontributif. Didapatkannya kombinasi semua fitur sebagai konfigurasi terbaik menandakan bahwa masalah pendeteksian kontroversi perlu didekati dari berbagai aspek.

Kata Kunci: *deteksi kontroversi, klasifikasi teks, pembelajaran mesin*

1. Introduction

The growing reach of Internet technology and the increasing of its usability has been able to bring the world mutually in a small country, where everyone is strongly connected to each other, just like one community. One can certainly mention incredible contributions of internet technology in many domains, such as education, research, public health, economics, entertainment, communication, journalism, etc. It is really clear how this technology has become one of the most important needs in our daily life.

In the area of journalism, Internet has provided many platforms that enables everyone to product and distribute reports on the interaction of events, facts, and ideas. We refer to this as citizen journalistic media, which is basically one of the forms of collaborative and social media. In the early stage of Internet, the one-way communication style of media has hindered citizen participation in terms of online journalistic activities. But, nowadays, the presence of social media, such as Weblogs, Microblogs, and Internet Forums has formed a new concept of communication, so called participatory journalism. Based on Bowman and Willis, [1], participatory journalism is defined as "the act of a citizen, or group of citizens, playing an active role in the process of collecting, reporting, analyzing and disseminating news and information". In addition to that, they also mentioned that "The intent of this participation is to provide indepen-

dent, reliable, accurate, wide-ranging and relevant information that a democracy requires." Differ from common professional journalism, participatory journalism requires no editorial oversight or formal journalistic workflow. Instead, it is the result of conversations in the online social media [1].

The opportunity to actively contribute in the participatory journalism has become great attractions for (non-professional) people. As a result, many online news websites have shifted towards facilitating a two-way communication platforms that enables citizen to share their journalistic writings within their websites. For example, one of the news agencies in the USA, CNN, has launched iReport 1. In Indonesia, there are several similar websites, such as Kompasiana 2 and Citizen 6 3.

Participatory journalistic media has several advantages as compared to common professional journalistic media, in the sense that the content of participatory journalistic media is actual and has more variation than common journalistic media. Unfortunately, it also has downsides since the creation process does not involve thorough editorial process. As a result, the accuracy of the content cannot be guaranteed. Moreover, the content has tendency to be biased, controversial, or provocative. This kind of readings can mislead many readers.

In our work, we focus on proposing a computational models to detect controversial articles due to its usefulness. Based on Merriam-Webster 4 dictionary, controversy is defined as "argument that involves many people who strongly disagree about something" or "strong disagreement about something among a large group of people". Controversial topics often involve many pros and cons around the topics. Wiley [2] mentioned that someone's understanding and stance on a particular controversial topic can be influenced by daily news or articles he consume everyday. Therefore, controversial topics should be carefully written. Unfortunately, readers usually do not realize that they are reading controversial articles. In addition to that, automatically recognizing controversial articles is not trivial. Knowing that a particular article is being controversial can help readers becoming more critical on information conveyed from the article.

One way to detect controversial articles is by looking at authorship debate occurred in the commentary section of the articles. However, when the total number of comments from an article reaches hundreds or thousands, it will be very difficult and tedious to manually read all comments and examine the quality of being controversial from the article. This problem motivates us to develop a computational model that can "automatically read" the article and its comments and determine its controversialness.

This paper is organized as follows. Section 2 describes related work on controversial article detection from the perspective of supervised approach. Then, section 3 presents the information of our annotated dataset for this task as well as our proposed approach. Section 4 discusses our experiment results. Finally, section 5 concludes our work and findings during the experiment.

Related Work

There have been several attempts in developing model for detecting controversial issues from several media, such as Wikipedia [3][4], News articles [5]–[7], and social media [8][9]. Popescu and Pennacchiotti [9] proposed a model for detecting controversial events from microblogs, like Twitter, related to some public figures in a fixed time period. They introduce the notion of twitter snapshot, i.e. a triple consisting of three concepts: target entity (e.g., Donald Trump), time period (e.g., one day), and a set of tweets talking about the target entity during the target time period. Their task is to assign a controversy score to each snapshot and rank the snapshots according to the controversy score. They argued that snapshot of controversial events provoke a public discussion, in which opposing opinions, surprise, or disbelief are easily found in the snapshot. Moreover, they coped with the problem using supervised machine learning models. Hence, several features were proposed to represent a snapshot, such as linguistic features, structural features, sentiment features, and "news buzz" features. They found that linguistic, structural, and sentiment features are highly ranked in terms of discriminative power.

Chimmalgi [6] focused on detecting controversial topics from social media such as comments and blogs, taking into account sentiment expressed in the comments, burstiness of comments, and controversy score. An annotated corpus consisting of 728 news articles was developed for training and evaluation purpose. Besides sentiment and structural features, Chimmalgi [6] also developed controversy-bearing term list from Wikipedia. Moreover, features derived from sentiment orientation score and controversy term list give the best discriminative power.

Allen et al., [10] conducted studies to detect disagree- ment in casual online forums such as Slashdot5. They presented a crowd-sourced annotated corpus for topic level disagreement detection. To develop the corpus, annotators were shown se-

1. http://ireport.cnn.com
2. http://www.kompasiana.com
3. http://citizen6.liputan6.com
4. http://www.meriam-webster.com/dictionary/controversy
5. https://slashdot.org/

veral topics and label them as containing disagreement or not. Furthermore, they found that disagreement detection is a subjective and difficult task since there are 22 topics (of 95 topics) has confidence scores below 55%. They formalized the problem as supervised learning using several hand-crafted features, such as rhetorical relations, sen-timent features, n-gram features, slashdot metafeatures, lexicon features, and structural features. Among those proposed features, the most discriminative features are those that include rhetorical information. Surprisingly, n-gram features harmed the model's performance.

Mejova et al. [7] studied the use of sentiment orientation information and biased words in 15 news portals. They showed empirical proof that controversial articles tend to reveal negative sentiment orientation and contain biased-words. Finally, Dori-Hacohen and Allan [11] proposed an automated approach to detect arbitrary webpages discussing controversial topics. Interestingly, they leveraged Wikipedia articles which bridge the gap between arbitrary webpages and rich metadata available in Wikipedia. They developed nearest neighbor classifier that maps webpages to the Wikipedia articles that discuss the same topic. Thus, the decision was solely based on those Wikipedia articles; if the Wikipedia articles are controversial, the corresponding webpages are also assumed to be controversial.

2. Methods

Data Collection

To build our dataset, we collected several articles from one of the famous participatory journalistic media in Indonesia, i.e. Kompasiana, during 9th - 10th April 2015. For our purpose, we selected those articles that have a considerable number of comments since analyzing comments is our main resource for feature extraction. In detail, first, we run a crawler to discover several popular topics. Moreover, we assumed that popular topics have been discussed in more than 100 articles. Second, for each popular topic, we run the second crawler to retrieve all related articles that have more than 60 comments. Next, after we collected a number of articles, we randomly selected around 500 articles from the collection for gold standard development. Finally, we manually annotated each of those articles as being controversy or non-controversy.

We obtained 304 articles as being controversy and 205 articles as non-controversy. Our controversial topics are mainly about political and law issue, such as Indonesian's presidential election in 2014 and Indonesian Corruption Eradication Commission. Table 1 shows the detail of our annotated corpus. We do not translate some terms since those terms are really specific to our domain.

TABLE 1
THE DETAIL OF OUR CONTROVERSY ARTICLE DATASET

Topic	Total	***Non-C**	***C**
pilpres 2014 (Presidential Election 2014)	203	96	107
pilkada jakarta (Jakarta Elecion)	47	31	16
KPK (Indonesian Corruption Eradication Commision)	37	22	15
pasangan capres (President Candidates)	29	12	17
jokowi capres 2014	22	8	14
jokowi nyapres	19	6	13
kawal pemilu	18	8	10
budi gunawan	15	9	6
debat capres (Election Debate) 2014	15	7	8
kompasiana baru	13	11	2

To create high quality corpus, we need to define annotation's guidelines, in which the definition of "being controversy" must be clear. Based on Indonesian's dictionary, controversy means something that sparks debate, while based on Merriam-Webster dictionary, controversy is defined as "argument that involves many people who strongly disagree about something". Hence, we formulated three questions that can help annotator to decide the label of an article: 1) Does the article form strong opinions toward a given issue or contain topics which are widely known as controversial (several topics are widely known as controversial, such as "gay", "atheism", "middle- east war", etc.), 2) Does the comment section of the article contain strong arguments or even aspersions, 3) Is the ratio between the number of pro's and con's comments of the article considerably balance.

Pre-Processing

Before we extracted the features of the articles, we followed several pre-processing steps: 1) we converted all the characters in the articles into their lowercased-version. 2) Explicit links were then removed from the article. 3) Non-canonical words were normalized into their canonical forms. For example, in Indonesian social media, we often see words, like "gak" and "nggak". Basically, they are non-canonical forms of the word "tidak", which means "not" in English. (4) Duplicate comments were then removed and considered as one distinct comment.

Proposed Methods and Features

We formulate our controversial article detection problem as a supervised classification problem us-

ing machine learning approach. Formally, given a set of label $L = \{\text{Controversy}, \text{NonControversy}\}$ and a set of articles $A = \{a_1, a_2, \ldots, a_n\}$, we seek a classifier function $F: A \rightarrow L$. Therefore, devising hand-crafted features is one of the most important steps in our task. In brief, we proposed 4 categories of features: META, STRUCTURAL, LINGUISTIC, and SENTIMENT features.

Meta Features (META)

This type of features leverage meta-data information found in the participatory journalistic media. There are two features belong to this category: the number of article views and the reader's evaluation.

1. The number of article views (META-1)
   Based on our observation, controversial article usually has a controversial title as well, such that many readers are interested in opening the article. As a result, the number of views upon a controversial article might have discriminative power in our classification task.
2. The reader's evaluation (META-2)
   In the participatory journalistic media, every reader can evaluate an article as being "inspirative", "interesting", "beneficial", or "actual". Therefore, an article is associated with 4 values denoting the number of votes of being "inspirative", "interesting", "beneficial", and "actual", respectively. We treat each value as a separate binary feature value.

Structural Features (STRUCTURAL)

This type of features captures the information regarding discussion activities that happened in the comment section of an article. Actually, the goal is that we need to know whether debate has occured in the comment section since debate is a good indicator for controversy.

1. Reply comments (STRUCTURAL-1)
   The first structural feature is the ratio between the number of reply comments and main comments. There are two types of comments, namely main comments and reply comments. Main comments directly response to the main article, while reply comments response to a particular main comment. We argue that when the number of reply comments is big enough, debates most likely occur in the comment section.
2. Distinct commentators (STRUCTURAL-2)
   We count the number of distinct commentators as our second structural feature. The rationale is that the more distinct commentators that a particular article has, the more intense the discussion among commentators.
3. Average cumber of comments (STRUCTURAL-3)
   This feature measures the activeness of a commentator. Someone who participate in a debate will most likely have more comments than those who don't.
4. Average length of comments (STRUCTURAL-4)
   Based on the observation conducted by Mishne and Glance [8], a comment that has good argument quality is quite lengthy. Therefore, we use the average length of comments as our structural feature. The length of a comment can be defined as the number of non-distinct words in the comment.
5. Maximum thread length (STRUCTURAL-5)
   Thread is identified as a single comment with its reply comments. Hence, thread length is defined as the number of reply comments in the thread itself. Based on our observation, comments that contain good arguments are usually replies to the a previous comment. In our work, we use the maximum thread length as our one of the structural features.

Linguistic Features (LINGUISTIC)

Based on our observation, we found that the controversialness of an article follows several linguistic patterns. Linguistic features mainly focuses on harnessing punctuations and lexical resource of bias words.

1. Question mark (LINGUISTIC-1)
   Question mark had been previously leveraged for the problem of controversial article detection [9][10]. Notice that in writing a sentence, question mark is frequently used to express curiosity or doubt about something. For example, in the following sentence, question mark can trigger debate on the commentary section.

   "oya, soal adipura, saya jamin 1000% hatta tahu. lho anak murid saya kelas 6 SD tahu semua. massa hatta enggak? jd nggak logis kalo hatta tdk tahu bedanya. massa anak SD lebih cerdas dari Hatta? nggak logis kan? hayolah akui itu." (Regarding adipura, i guarantee 1000% that Hatta knows this issue. All my elementary students know this issue very well, why not with Hatta? Are my elementary students smarter than Hatta? it really doesn't make sense, does it?)

   Specifically, we use the number of comments containing at least one question mark and the number of question marks in the main article body as our features.

2. Exclamation mark (LINGUISTIC-2)
   Just like question mark, exclamation mark was also used by Allen et al. [10] as one of their features for the same task. Exclamation mark usually indicates spirit and anger in the journalistic articles, in which this usually happens in a debate situation. We use the number of comments containing at least one exclamation mark and the number of exclamation marks in the main article body as our features.
3. Capital letters (LINGUISTIC-3)
   In a debate situation, writers usually express their arguments using capital letters to emphasize a certain issue or topic. We use the number of comments containing at least three words with capital letters. In addition, we use th-ree as our threshold number to compensate abbreviations, in which all letters are capitalized.
4. Bias lexicon (LINGUISTIC-4)
   Bias words tend to support a certain opinion, which means that they are not neutral. Furthermore, Mejova et al. (2014) found that controversial articles contain many bias words. To create Indonesian bias lexicon, first, we automatically translated 654 English bias words developed by Recansens et al. [4]. After that, we checked the translated bias word manually and discovered 602 bias words ready to use. The following list shows several words from our collection of 602 bias words. For main article body and commentary section, we use the proportion of bias words in the document as our feature.
5. Example of bias words
   aborsi (abortion), *fanatisme* (fanatism), *genosida* (genocide), *bom* (bomb), *homoseksual* (homosexual), *minoritas* (minority), *terorisme* (terorism), *skandal* (scandal), *sosialis* (socialism), *komunis* (communism), *muslim* (moslem), *katolik* (catholic), *rasis* (racism), *revolusi* (revolution), *yahudi* (jew), *zionis* (zionism), *propaganda* (propa-ganda)

Sentiment Features (SENTIMENT)

This type of features leverages subjectivity in the body and commentary section of an article. In our case, we hypothesize that a controversial article tends to reveal pro and contra about a particular topic. In addition, pro and contra toward a particular topic are usually expressed using opinionated words. As a result, we can employ several techniques for detecting sentiment orientation inside the articles and their comments. Furthermore, our techniques harness Indonesian sentiment lexicon developed by Vania [12], which contains 416 positive words and 581 negative words.

1. Sentiment score (SENTIMENT-1)
   Suppose Pos denotes the number of positive words, Neg denotes the number of negative words, and Neu denotes the number of neutral words in a document (body of articles or commentary section), we compute the sentiment score of the document as follow.

$$\text{sentiment} = \frac{\text{Pos} - \text{Neg}}{\text{Pos} + \text{Neg} + \text{Neu}} \quad (1)$$

   Equation(1) yields a value in range $[-1, +1]$, where value less than zero means that the overall document reveals negative sentiment, and vice versa. From this score, we devise three feature values: (1) sentiment score of the article body, (2) average sentiment scores of all comments (commentary section can have more than one comment), the difference between article body's sentiment score and average sentiment scores of all comments. The last feature is proposed since, based on our observation, controversialness of an article is often triggered by the opinion clashes or differences that happen between the content of article body and commentary section.

2. Standard deviation of sentiment score (SENTIMENT-2)
   This feature is based on rationale that the more controversial an article is, the more various the sentiment expressed inside the commentary section. The variance of all sentiment scores inside the comments can be captured using the following standard deviation formula.

$$\text{std} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - \mu_s)^2} \quad (2)$$

   where N is the number of comments, S_i is the sentiment score of i-th comment, and μ_s is the average of all comments' sentiment scores.

3. Mixed sentiment score (SENTIMENT-3)
   We also use mixed sentiment score proposed by Popescu and Pennacchiotti [9]. In their original paper, they use this scoring mechanism to detect controversialness on microblogs, such as tweets. In our case, instead of tweets, we apply this scoring formula to all comments.

$$\text{mixSentiment} = \frac{\min(\text{Pos}, \text{Neg})}{\max(\text{Pos}, \text{Neg})} \frac{\text{Pos} + \text{Neg}}{\text{Pos} + \text{Neg} + \text{Neu}} \quad (3)$$

4. Ratio of positive and negative comment (SENTIMENT-4)
   For the last feature, we use the ratio value of positive and negative comments. The ratio of positive and negative comments is determined as Pos/N and Neg/N, respectively, where N is the total number of comments, and Pos and Neg are the number positive and negative comments, respectively.

3. Results and Analysis

Metrics and Experiment Settings

To evaluate the performance of our proposed model, we use precision, recall, and F1-score as our evaluation metrics. We also use 10-fold cross validation and employ Logistic Regression and Support Vector Machine as our classifiers. Furthermore, before we did experiments, we made our dataset balance by employing oversampling technique, namely SMOTE [13]. Table 2 shows the comparison of our dataset before and after oversampling.

After that, we performed several experiments to see the contribution of each feature type, as well as find the best combination of features for our classification model. First, we tried each feature group separately. Second, we performed feature ablation study to see the contribution of each feature group relative to the others. Finally, we compared the effect of two feature sources, i.e., the body of article and the commentary section, upon classification performance to see which source of information contribute most to the task.

Result

First, we run Chi-Square test to see the contribution of every single feature for the classification task. The result can be seen in Table 3. Moreover, we only show top-10 most discriminative features. It is interesting to see that all feature groups have their representatives in top-10, except for META features. Next, to see the effect of each feature group as a whole, we conducted experiment involving only one feature group. As can be seen in Table 4, META features gives the worst performance compared to the other feature groups (around 61.5% and 59.2% for LogReg and SVM, respectively). This result is actually inline with the information described in Table 3. STRUCTURAL and LINGUISTIC feature groups are considerably important our detecting controversial contents. When we only used STRUCTURAL feature group, the accuracy achieved 79.4% with Logistic Regression model.

Next, we performed feature ablation study (Table 5), i.e., empirical analysis task that explores the contribution of each feature group by omitting each group while keeping the other feature groups. As can be seen in Table 5, the worst accuracy was yielded when we omit STRUCTURAL feature group, which means that STRUCTURAL feature group is the most discriminative feature group for this task. Until now, it seems like Logistic Regression model outperforms Support Vector Machine. Finally, we performed an experiment to see which source of features (either from article body or comment section) has the most discrim- inative power. Table 6 shows that information (features) from commentary section has notable contribution for our classification task. When we used features extracted from commentary section, the performance reached more than 81%, while the features extracted from article body resulted in much lower classification performance (below 63% in terms of accuracy). The best performance in our experiment was achieved when we used all features and Logistic Regression as our classifier, i.e., 82.8% and 83.1% in terms of accuracy and F1-score, respectively. Moreover, the value of precision and recall for "Controversy" label tend to be similar in many scenarios.

It is also worth to know several reasons for False Positive in our classification task. False Positive means that we mis-classify non-controversy article as controversy article. Based on our observation, this is mostly due to lengthy SPAM comments which can contain around 1180 words. This harms the performance of our classifie by reducing the discriminative power of some features, including STRUCTURAL-4 feature. The other case is due to lack of opinion lexical resources (for In-

TABLE 2
OUR DATASET (BEFORE AND AFTER OVERSAMPLING)

Condition	Non-Controversy	Controversy	Total
Original	304	205	509
Oversampled (SMOTE)	304	304	608

TABLE 3
CHI-SQUARE VALUE OF EVERY SINGLE PROPOSED FEATURE

Rank	Feature Name	Category Name
1	Average Length of Comments	STRUCTURAL-4
2	Capital Letters	LINGUISTIC-3
3	The Ratio of Negative Comments	SENTIMENT-4
4	Mixed Sentiment Score	SENTIMENT-3
5	Maximum Thread Length	STRUCTURAL-5
6	Bias Lexicon	LINGUISTIC-4
7	Question Mark	LINGUISTIC-1
8	Average Sentiment Scores in Comments	SENTIMENT-1
9	Average Number of Comments	STRUCTURAL-3
10	Reply Comments	STRUCTURAL-1

TABLE 4
THE PERFORMANCE OF DETECTION - USING ONLY ONE FEATURE GROUP

Feature Group	Class	Logistic Reg				SVM			
		Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
META	Non-Controversy	0.61	0.60	0.61	0.61	0.65	0.38	0.48	0.59
	Controversy	0.61	0.62	0.61		0.56	0.80	0.66	
STRUCTURAL	Non-Controversy	0.77	0.83	0.80	**0.79**	0.67	0.89	0.76	0.73
	Controversy	**0.82**	**0.75**	**0.78**		0.84	0.56	0.67	
LINGUISTIC	Non-Controversy	0.76	0.77	0.76	0.76	0.74	0.79	0.76	0.76
	Controversy	0.76	0.76	0.76		0.78	0.72	0.75	
SENTIMENT	Non-Controversy	0.72	0.71	0.71	0.72	0.75	0.62	0.68	0.71
	Controversy	0.71	0.73	0.72		0.68	0.79	0.73	

TABLE 5
THE PERFORMANCE OF DETECTION – FEATURE ABLATION STUDY

Feature Group (without)	Class	Logistic Reg				SVM			
		Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
META	Non-Controversy	0.80	0.80	0.80	0.80	0.80	0.75	0.78	0.78
	Controversy	0.80	0.80	0.80		0.77	0.81	0.79	
STRUCTURAL	Non-Controversy	0.78	0.78	0.78	0.78	0.80	0.72	0.76	0.77
	Controversy	0.78	0.78	0.78		0.74	0.82	0.78	
LINGUISTIC	Non-Controversy	0.78	0.82	0.80	0.80	0.76	0.74	0.75	0.75
	Controversy	0.81	0.78	0.79		0.75	0.77	0.76	
SENTIMENT	Non-Controversy	0.80	0.83	0.82	**0.81**	0.72	0.88	0.79	0.77
	Controversy	**0.82**	**0.80**	**0.81**		0.84	0.66	0.74	

TABLE 6
THE PERFORMANCE OF DETECTION – THE CONTRIBUTION OF EACH SOURCE OF FEATURES

Feature (Classified based on Source)	Class	Logistic Reg			
		Prec	Rec	F1	Acc
META	Non-Controversy	0.80	0.80	0.80	0.80
	Controversy	0.80	0.80	0.80	
STRUCTURAL	Non-Controversy	0.78	0.78	0.78	0.78
	Controversy	0.78	0.78	0.78	
LINGUISTIC	Non-Controversy	0.78	0.82	0.80	0.80
	Controversy	0.81	0.78	0.79	
SENTIMENT	Non-Controversy	0.80	0.83	0.82	**0.81**
	Controversy	**0.82**	**0.80**	**0.81**	

donesian Language) such that some opinionated words are misclassified in terms of its polarity. Moreover, this problem also raises because our sentiment analysis model does not have capability in detecting target of the sentiment evaluation. For example, the following two sentences have different polarity in the context of Indonesian Presidential Election in 2014, yet they actually support each other towards one candidate.

- *kita yg cerdas sudah pasti pilih no. 2* (we, smart people, will absolutely choose number 2)
- *sudah jelas gak bisa memenuhi janji pertama lalu membuat janji yg lebih tidak masuk akal lagi. cepek deh payah banget.* (It is clear that he was not able to fulfill his first promise, yet he made another new promise which doesn't make sense. how loser you are)

There were two candidates at that time. The first sentence evaluates the first candidate as positive, while the second one evaluates the second candidate as negative, which means that these two sentences actually support the first candidate. As a result, incapability to detect opinion target would certainly drop the performance. We also argue that this phenomenon is the reason why our proposed SENTIMENT feature group is not really superior in our case. In addition, we are not really interested in False Negative since we focus on precision, instead of recall, for this controversial detection task.

4. Conclusion

In this paper, we have shown our approach to automatically detect controversial articles due to several motivations. We proposed a supervised machine learning approach harnessing several handcrafted features. Furthermore, our work mostly focus on feature engineering, in which there are four main feature groups: META, STRUCTURAL, LINGUISTIC, and SENTIMENT feature groups. To see the contribution of every single feature and each feature group, we performed several experiments, including feature ablation study and feature ranking (based on discriminative power). We have found that STRUCTURAL and LINGUISTIC feature groups contribute the most to the classification task, while META feature group seems not to be really important for the task. For SENTIMENT

feature group, we argue that its contribution is considerably low due to the fact that opinion expressed in the sentence is quite complex. As a result, more advanced aspect-based sentiment analysis task is required to improve the performance of the task.

The best performance so far is achieved when we use all proposed feature with Logistic Regression as our model (82.89% in terms of accuracy). Finally, we also conducted experiment to see which source of features (either comments or article body) that contributes the most to the classifier's performance. The result show that features extracted from article body seem not to be really discriminative, compared to features extracted from commentary section. When we used all features extracted from article body, the performance achieved 62.9% in terms of accuracy. On the other hand, features from commentary section can result in 81.7% in terms of accuracy. The fact that the best accuracy is yielded by using all features indicates that the task of controversy detection has to consider many aspects of the article.

In the future, we plan to collect more dataset covering more controversial topics. When our corpus is large enough, we can apply state-of-the-art deep learning approach for text classification, in which feature engineering is no longer needed. In fact, our main reason why we employ feature engineering approach is due to small dataset size. However, our hand-crafted features are really important in under- standing hidden information gleaned inside a controversial articles. We can somehow combine the information from our proposed features and automatic learned-features inferred by deep learning model to achieve better performance.

References

[1] S. Bowman and C. Willis. (2003) We media. the media center.

[2] J. Wiley, "A fair and balanced look at the news: What affects memory for controversial arguments?" *Journal of Memory and Language*, vol. 53, no. 1, pp. 95–109, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0749596X05000215

[3] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi, "He says, she says: conflict and coordination in wikipedia," *In Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 453– 462, 2007.

[4] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language," in *ACL*, 2013.

[5] Y. Choi, Y. Jung, and S.-H. Myaeng, "Identifying controversial issues and their subtopics in news articles," in *Proceedings of the 2010 Pacific Asia Conference on Intelligence and Security Informatics*, ser. PAISI'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 140–153. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13601-6 16

[6] R. V. Chimmalgi. (2013) Controversy detection in social media. masters thesis. louisiana state university.

[7] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo, "Con- troversy and sentiment in online news," *CoRR*, vol. abs/1409.8152, 2014. [Online]. Available: http://arxiv.org/abs/1409.8152

[8] G. Mishne and N. Glance, "Leave a reply: An analysis of weblog comments," in *In Third annual workshop on the Weblogging ecosystem*, 2006.

[9] A.-M. Popescu and M. Pennacchiotti, "Detecting controversial events from twitter," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 1873–1876. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871751

[10] K. Allen, G. Carenini, and R. Ng, "Detecting disagreement in conversations using pseudo-monologic rhetorical structure," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 1169–1180. [Online]. Available: http://www.aclweb.org/anthology/D14-1124

[11] S. Dori-Hacohen and J. Allan, *Automated Controversy Detection on the Web*. Cham: Springer International Publishing, 2015, pp. 423–434. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16354-3 46

[12] C. Vania, "Perolehan opini pada dokumen blog berbahasa indonesia (opinion retrieval on indonesian weblogs)," *Bachelors thesis, Universitas Indonesia*, 2009.

[13] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *CoRR*, vol. abs/1106.1813, 2011. [Online]. Available: http://arxiv.org/abs/1106. 1813