# Towards Argumentative Opinion Mining in Online Discussions

Jérémie Clos
Robert Gordon University
j.clos@rgu.ac.uk

Nirmalie Wiratunga
Robert Gordon University
n.wiratunga@rgu.ac.uk

Joemon Jose
University of Glasgow
joemon.jose@glasgow.ac.uk

Stewart Massie
Robert Gordon University
s.massie@rgu.ac.uk

Guillaume Cabanac
University of Toulouse
guillaume.cabanac@univ-tlse3.fr

## 1 Introduction

Online discussion forums (Figure 1) typically manifest into tree-like structures that are reminiscent of argument trees. Whilst these discussion forums contain a wealth of information related to people's opinions they also include implicit argumentation information. However unlike argument trees any relationship between posts in a discussion tree remains implicit. In recent years there has been considerable interest in harnessing opinionated knowledge that is buried in discussion forums in a variety of domains (e.g. retail, health and politics). However to-date the main focus of opinion mining has been on generating and applying sentiment lexicons or developing supervised learning algorithms that pay little to no regard to the underlying role of implicit information contained in the conversation between users.

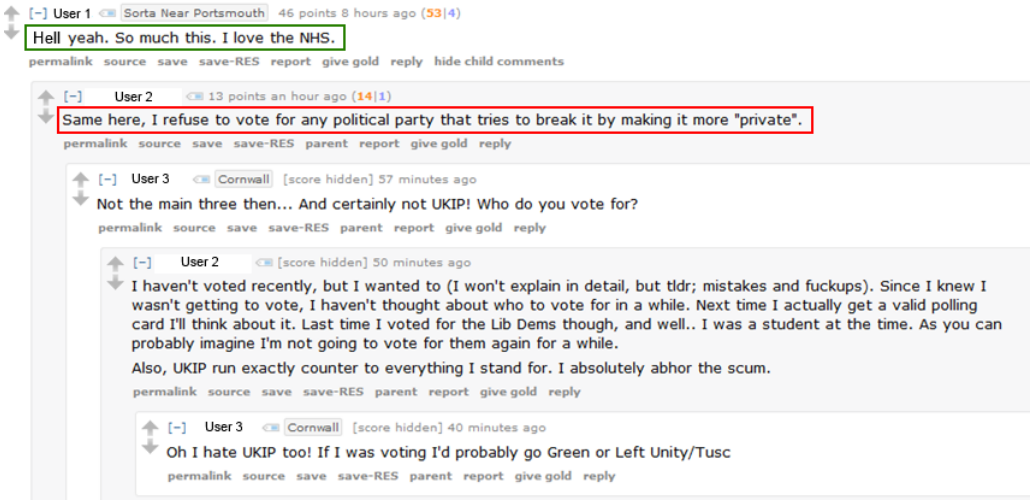


Figure 1: Part of a discussion tree from Reddit where users discuss United Kingdom politics.

Consider the discussion excerpt presented in figure 1. We consider the outcome of applying the SMARTSA state-of-the-art sentiment analysis system [MWLG13] to argumentative content. The aggregated sentiment of comment 1 (circled in green) is positive with a value of $0.18$, while the aggregated sentiment of comment 2 (circled in red) is negative with a value of $-0.23$. A casual observation suggests that the opinion expressed in both comments are not opposed but aligned. A close examination shows us that there is indeed a discrepancy between the actual opinions expressed, which are similar, and what a sentiment analysis algorithm tells us, which is that comment 1 is positive while comment 2 is negative. This is due to the fact that sentiment analysis algorithms such as SMARTSA generally rely only on the aggregated sentiment mined from the terms contained in the sentences, considering user comments in a vacuum. We tackle this issue by exploring the use of argumentation constructs for argumentative opinion analysis in social media.

Abstract or bipolar argumentation frameworks [Dun95, ACLSL08] can be seen as graphical models representing the argumentation process. They represent arguments as nodes and relations between arguments in the form of directed edges, which can be either untyped (Dung argumentation frameworks [Dun95]) or typed (Bipolar argumentation frameworks [ACLSL08]). These high level graphical frameworks are generally used to represent the argumentation process between multiple agents, but are less applicable to the analysis of such content because of its informal and noisy nature. Accordingly, there is a need for novel frameworks better able to perform more flexible analysis. We formulate in our work the task of argumentative opinion analysis as the task of identifying the relationships between two user-generated

comments as a binary classification problem. For practical purposes and computational efficiency, we simplify the traditional argumentation theoretic concepts of argument defeat and support into opinion disagreement and agreement. We assume any comment to be either in agreement or disagreement with its parent comment, with no partial membership. Let $C$ be the domain of user-generated comments, $\Re_D$ and $\Re_A$ the disagreement and agreement relations such that $\Re_D(a, b)$ means "*comment a disagrees with comment b*" and $\Re_A(a, b)$ "*comment a agrees (or does not disagree) with comment b*" where $a \in C$ and $b \in C$. Disagreement is defined as a coarser and symmetric relation derived from the typical defeat (rebuttal/undercut) relation. We use a weak definition of agreement which includes both explicit agreement and absence of explicit disagreement. This is done so as to encompass comments with an absence of argumentative relationship with their parent comment. Accordingly the goal of argumentative opinion classification is the detection of the type of argumentative relation which links a comment to its parent comment.

In this work we focus more precisely on the classification of the child post and conduct a comparative study to establish the transferability of standard text representation strategies from information retrieval for argumentative opinion classification. We introduce a novel term weighting strategy, $TF_w$ that is biased by argumentation constructs present in a document. Our promising initial results show $TF_w$ to be very sensitive to the post length and specifically suited to longer comments.

# 2 Background

The domain of text classification, illustrated in figure 2, has been extensively explored in the text mining community. The process of text mining and the learning of a model, begins with the representation of the textual content in a way which allows classifiers to learn how to differentiate instances from different classes. One common way to do so is by employing the bag-of-words (BOW) representation. In this representation scheme any text is represented by a vector of the terms it contains, each term being either binary (e.g. present or absent) or weighted (e.g. by the number of times it appears in the text). The union of all terms induces a n-dimensional vector space[Sal79] (where n is the number of unique terms in the corpus) on which each document is projected and on which a classifying model can be learned. When computing this projection, both local and global weighting schemes are used to determine the value of each component of the vector representation by computing respectively the local (in the document) and global (in the corpus) importance of the term associated to that component[SJ72]. Because of the small size of our documents, we ignore at first the contribution of global weighting schemes and put it aside for later study. We use the following local weighting schemes for our text representation:

- Binary: $Bin$ is set to 1 if the term is present in the document and 0 otherwise ; for a term $t$ and a document $d$:

$$Bin(t, d) = \{0, 1\} \tag{1}$$

- Raw term frequency: $TF_r$ is set to the number of occurrences of a term in the text ; for a term $t$ and a document $d$:

$$TF_r(t, d) = freq(t, d) \tag{2}$$

- Smoothed frequency: $TF_s$ is set to the number of occurrences of a term in the text, smoothed on a logarithmic scale to avoid having frequencies in different orders of magnitude ; for a term $t$ and a document $d$:

$$TF_s(t, d) = log(1 + freq(t, d)) \tag{3}$$

We address in this work the effectiveness of each of these representation schemes for the task of classifying disagreements and agreements, and propose directions to overcome their potential weaknesses when dealing with comments of varying length. The binary representation ($Bin$) is well known for its effectiveness against the classification of microblogs and texts of small size but is less scalable to larger documents. On the other hand, the raw term frequency ($TF_r$) is known to be effective for longer documents but is not equipped to deal with documents of varying length: extremely long documents can overpower smaller ones by the sheer number of terms used in them. Finally, smoothed term frequency ($TF_s$) helps alleviate that weakness by projecting that frequency on a logarithmic scale, which reduces the gap between short and long documents and usually performs the best on heterogeneous collections of documents.

# 3 Contribution

## 3.1 Lexicon-augmented Term Frequency Modelling

As part of the learning process we can integrate background knowledge in the form of a lexicon (represented in grey in figure 2). A lexicon is used to capture domain knowledge such as the association of certain terms to a sentiment polarity (in lexicon-based sentiment analysis algorithms such as [MWLG13]) or to associate certain linguistic patterns to a degree

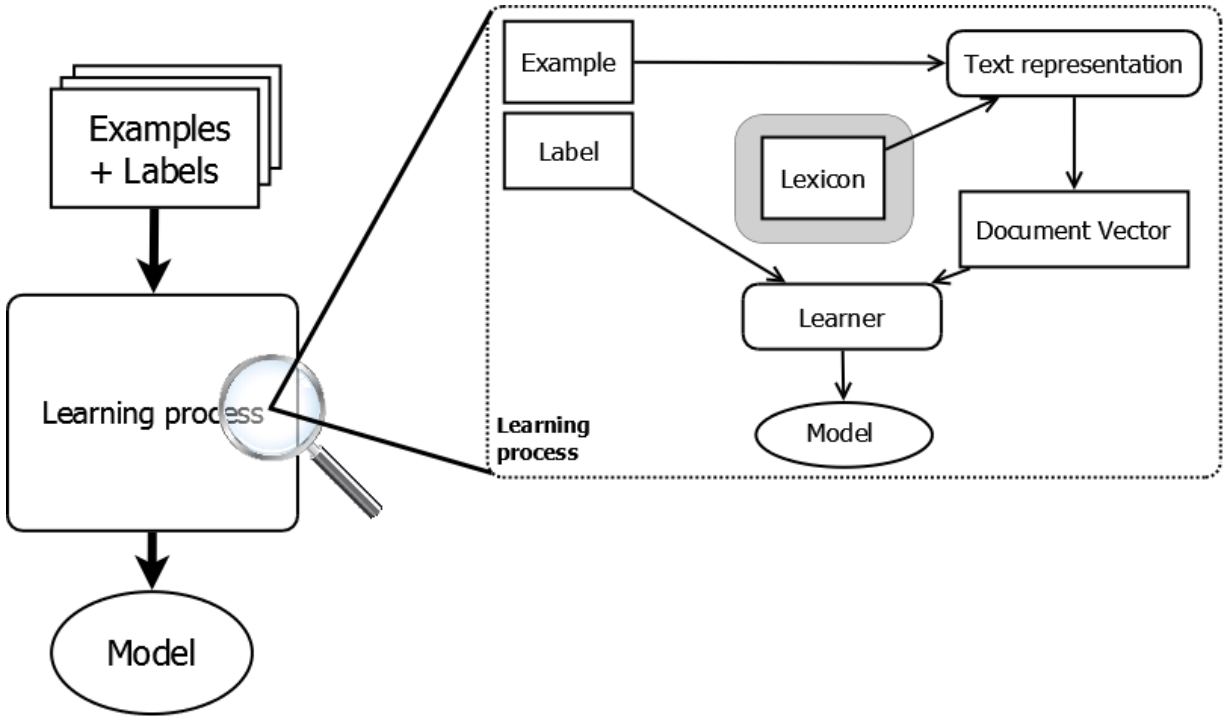


Figure 2: The learning phase of a text mining algorithm.
Examples labeled with their class are given as input and processed into document vectors and then sent along with their label to the inductive learner which outputs a model of the data.

of argumentativeness, such as in our approach. There are two typical strategies that can be adopted to generate lexicon guided representations: as a way to derive features which are then used to train a classifier; or as a way to derive features which are then used to directly classify, such as by aggregating sentiment values in lexicon-based sentiment analysis. We employ the former whereby argumentation constructs from a lexicon are used to focus the vector space projection of comments.

Weighted local term frequency, $TF_w$, is a local weighting scheme that assigns higher weights to terms that frequently occur in sentences that are rich in argumentation constructs. Its goal is to improve classification of larger comments by increasing the influence of sentences that are known to contain patterns from a lexicon that are potential signals for the presence of arguments. Given a document $d$ we can formulate the weight for a term, $t \in d$ as follows:

$$TF_w(t,d) = log\left(1 + \sum_{s \in d} \sum_{t \in s} w(s)\right) \tag{4}$$

The weight $w(s)$ of each sentence $s$ is computed using a lexicon of argumentation structures (see [SRW07]) which act as evidence of presence of reasoning patterns in the sentences containing them. A uniform weight is assigned to all structures and the weight of a sentence is calculated as follows, with $a$ being argument structures present in sentence $s$:

$$w(s) = 0.2 + log\left(1 + \sum_{a \in s} w(a)\right) \tag{5}$$

## 3.2 Preliminary experiments

### 3.2.1 Experimental setup

We experiment using our three baselines ($Bin$, $TF_r$ and $TF_s$ for binary, raw frequency and smoothed frequency) and the $TF_w$ weighting scheme with a novel dataset extracted from Reddit[1] for this specific task. The data is labelled using a noisy labelling approach inspired from distant supervision learning [MBSJ09] whereby highly discriminative expressions such as "I agree" and "you are wrong" are used as cues to class labels agreement and disagreement.

Results are computed with an artificially balanced dataset (to avoid majority class bias) and compared using their classification accuracy, which is computed as follows:

[1]Reddit is a social website in which users can register to post stories and comment or vote on other stories or comments. It can be accessed at `http://www.reddit.com`

$$\text{accuracy} = \frac{|\text{true disagreements}| + |\text{true agreements}|}{|\text{true disagreements}| + |\text{true agreements}| + |\text{false agreements}| + |\text{false disagreements}|} \quad (6)$$

In this equation true disagreements and true agreements are the instances correctly classified as disagreements and agreements, and conversely false disagreements and false agreements are the instances incorrectly classified as disagreements and agreements.

### 3.2.2 Results and discussion

Figure 3 presents accuracy results for increasing document length. In general the utility of $TF_w$ increases with increasing document length sizes improving over other weighting schemes once length is greater than 10. This is not suprising since the opportunity to discriminate between sentences that do and do not contain argumentative constructs improves with larger comments. In fact we find that on average 17 patterns from our lexicon are triggered as opposed to an average of 3 for comments with less than 10 sentences.

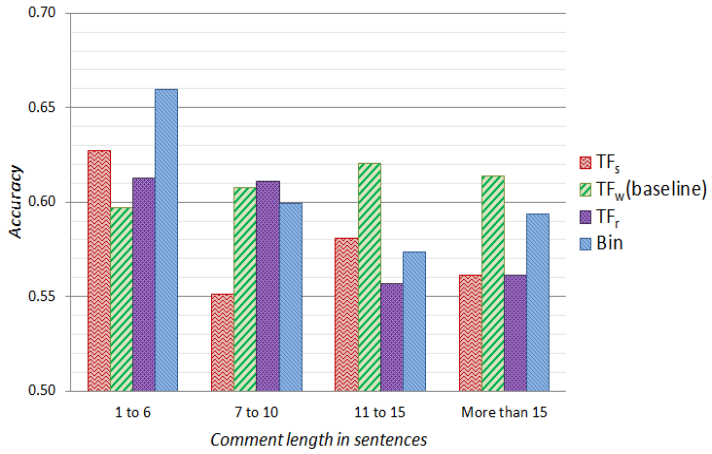


Figure 3: Comparative analysis of representations.

| # sentences | Under 10 | Over 10 |
|---|---|---|
| # comments | 40,206 | 6,189 |
| Avg terms/sentence | 16 | 19 |
| Avg sent/comment | 4 | 22 |
| Avg triggered patterns | 3 | 17 |
| # terms | 62,860 | 57,098 |
| Common vocabulary | 29,069 terms | |
| Total vocabulary | 90,889 terms | |

Table 1: Statistics about our dataset.

Our results warrant further investigation into other weighting schemes (for example based on language modeling as described in [PT10] or supervised term weighting approaches such as described in [DLY14]) and how to use them to improve our classification accuracy. Additional work in the improvement of the $TF_w$ weighting scheme (e.g., the development of an improved lexicon or the refinement of the weighting mechanism so as to improve accuracy on short comments) is also needed for the advancement on the task of argumentation opinion classification. Indeed, since the overwhelming majority of comments in social media are composed of 4 sentences on average (see table 1 for descriptive statistics around the 10 sentences pivot point), specialized term weighting schemes such as $TF_w$ do not hold up against others.

A complementary analysis on specialised sub-corpora (in the following domains: general debates, news and political discussions) also showed us that the behaviour of the representation schemes seems to be topic-dependent: none of the sub-topics analysed shows a behaviour similar to the entire corpus, $TF_w$ faring consistently worse than the traditional information retrieval term weighting schemes with an accuracy on average $4.8\%$ under the best baseline. However, given the relatively small size of our subcorpora (1,500 instances per corpus), we cannot hold any conclusion as to the scalability of such results to bigger datasets.

# 4 Conclusion and future work

We introduced in this position paper the task of "argumentative opinion classification" as the classification into agreement or disagreement of a user-generated comment relative to its parent comment. We introduced a local term weighting strategy that is biased in favour of the presence of argumentative constructs in sentences. We conduced an experiment using an automatically labelled dataset extracted from Reddit and conclude the necessity of further investigation in finer-grained weighting schemes (such as language model-based weighting), the use of supervised term weighting approaches and the improvement of $TF_w$ via the learning of a better lexicon. Further work will also include the scaling of our experiment to a larger dataset and the creation of a standard corpus for argumentative opinion classification in social media.

# References


[ACLSL08] Leila Amgoud, Claudette Cayrol, Marie Christine Lagasquie-Schiex, and Pierre Livet. On bipolarity in argumentation frameworks. *Int. J. Intell. Syst.*, 23(10):1062–1093, 2008.

[DLY14] Zhi-Hong Deng, Kun-Hu Luo, and Hong-Liang Yu. A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, 41(7):3506 – 3513, 2014.

[Dun95] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[MBSJ09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[MWLG13] Aminu Muhammad, Nirmalie Wiratunga, Robert Lothian, and Richard Glassey. Domain-based lexicon enhancement for sentiment analysis. In *SMA BCS-SGAI*, pages 7–18, 2013.

[PT10] Georgios Paltoglou and Mike Thelwall. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1386–1395, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Sal79] Gerard Salton. Mathematics and information retrieval. *Journal of Documentation*, 35(1):1–29, 1979.

[SJ72] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[SRW07] Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6, 2007.