From Argumentation Mining to Stance Classification

Parinaz Sobhani[1], **Diana Inkpen**[1], **Stan Matwin**[2]
[1] School of Electrical Engineering and Computer Science, University of Ottawa
[2] Faculty of Computer Science, Dalhousie University
[2] Institute of Computer Science, Polish Academy of Sciences
psobh090@uottawa.ca, diana.Inkpen@uottawa.ca,stan@cs.dal.ca

Abstract

Argumentation mining and stance classification were recently introduced as interesting tasks in text mining. In this paper, a novel framework for argument tagging based on topic modeling is proposed. Unlike other machine learning approaches for argument tagging which often require large set of labeled data, the proposed model is minimally supervised and merely a one-to-one mapping between the pre-defined argument set and the extracted topics is required. These extracted arguments are subsequently exploited for stance classification. Additionally, a manually-annotated corpus for stance classification and argument tagging of online news comments is introduced and made available. Experiments on our collected corpus demonstrate the benefits of using topic-modeling for argument tagging. We show that using Non-Negative Matrix Factorization instead of Latent Dirichlet Allocation achieves better results for argument classification, close to the results of a supervised classifier. Furthermore, the statistical model that leverages automatically-extracted arguments as features for stance classification shows promising results.

1 Introduction

In the past, people were only the consumers of information on the web. With the advent of Web 2.0, new tools for producing User Generated Content (UGC) were provided. Consequently, huge amounts of text data is generate every day on the web. As the volume of this unstructured data increases, the request for automatically processing UGC grows significantly. Moreover, this new source of information and opinions contains valuable feedback about products, services, policies, and news and can play an important role in decision making for marketers, politicians, policy makers and even for ordinary people.

So far, there has been a great effort toward subjectivity analysis of sentiment and opinion mining of reviews on concrete entities such as product or movies (Pang et al., 2002), (Dave et al., 2003), (Pang and Lee, 2005); however, this line of research does not fit online discussions opinion mining where comments not only contain the sentiment/stance of the commenter toward the target, but also convey personal beliefs about what is true or what action should be taken. This kind of subjectivity is called argumentation (Wilson and Wiebe, 2005). Argumentation analysis is more focused on the reason for author's overall position.

Stance has been defined as the overall position toward an idea, object or proposition (Somasundaran and Wiebe, 2010). There has been growing interest in stance classification particularly for online debates (Walker et al., 2012a), (Hasan and Ng, 2013). To the best of our knowledge, our paper is the first work for stance classification of the news comments considering particular news as target to investigate the overall position toward it.

Argument tagging was first introduced as a task in (Boltuzic and Šnajder, 2014) in which the arguments were identified from a domain-dependent predefined list of arguments. An argument tag is a controversial aspect in the domain that is abstracted by a representative phrase/sentence (Conrad et al., 2012).

In our paper, a new framework for argument tag-

Proceedings of the 2nd Workshop on Argumentation Mining, pages 67–77,
Denver, Colorado, June 4, 2015.

ging at document-level based on topic modeling, mainly Non-Negative Matrix Factorization, is proposed. The main advantage of this framework is that it is minimally supervised and no labeled data is required.

The correlation between stance labels and argument tags has been addressed in different studies (Boltuzic and Šnajder, 2014) (Hasan and Ng, 2014). In our research, a statistical model for stance classification based on the extracted arguments is suggested, while in previous research stance labels were exploited for argument tagging.

Nowadays, several popular news websites like CNN and BBC allow their readers to express their opinion by commenting; these kinds of commentspheres can be considered as type of social media. Consequently, visualizing and summarizing the content of these data can play a significant role in public opinion mining and decision making. Considering the huge volume of the news comments that are generated every day, manual analysis of these data may be unfeasible. In our research, a corpus of news comments is collected and annotated and is made available to be deployed as a benchmark in this field [1]. Hence, it provides opportunities to further investigate automatic analysis of such types of UGC.

2 Related Work

In (Somasundaran et al., 2007), two types of opinions are considered: sentiment and arguments. While sentiment mainly includes emotions, evaluations, feelings and stances, arguments are focused on convictions and persuasion.

Stance Classification One of the first works related to stance classification is perspective identification (Lin et al., 2006), where this task was defined as subjective evaluation of points of view. Supervised learning has been used in almost all of the current approaches for stance classification, in which a large set of data has been collected and annotated in order to be used as training data for classifiers. In (Somasundaran and Wiebe, 2010), a lexicon for detecting argument trigger expressions was created and subsequently leveraged to identify arguments. These extracted arguments together with sentiment expressions and their targets were employed in a supervised learner as features for stance classification. In (Anand et al., 2011), several features were deployed in their rule-based classifier, such as unigrams, bigrams, punctuation marks, syntactic dependencies and the dialogic structure of the posts. The dialogic relations of agreement and disagreements between posts were exploited in (Walker et al., 2012b),(Ghosh et al., 2014), likewise; while in this paper our aim is to investigate stance without considering the conversational structure which is not always available.

Argument Tagging In (Albert et al., 2011), argument mining for reviews was introduced in order to extract the reasons for positive or negative opinions. Argumentation analysis can be applied at different text granularities. In (Conrad et al., 2012), a model for argument detection and tagging at sentence-level was suggested. In our research, argument tags were organized in a hierarchical structure inspired by a related field in political science "Arguing Dimension" (Baumgartner et al., 2008). In (Hasan and Ng, 2014), a reason classifier for online ideological debates is proposed. In this method document-level reason classification is leveraged by aggregating all sentence-level reasons of a post. Our proposed method tags arguments at document-level and unlike previous works is minimally supervised.

Topic Modeling Topic modeling in more informal documents is more challenging due to the less organized and unedited style of these documents. Topic-modeling has been used in sentimental analysis and opinion mining to simultaneous investigate the topics and the sentiments in a text (Titov and McDonald, 2008a), (Mei et al., 2007). One of the most popular approaches for topic modeling is Latent Dirichlet allocation (LDA) (Blei et al., 2003). This probabilistic model has been extended in (Titov and McDonald, 2008b) to jointly model sentiments and topics in an unsupervised approach. LDA topic modeling was also employed for automatic identification of argument structure in formal documents of 19th century philosophical texts (Lawrence et al., 2014). LDA was applied on the target corpus and the resulting topics were exploited to find similarities between the different propositions. Non-Negative Matrix Factorization (NMF) (Lee and Seung, 2001)

[1] https://github.com/parinaz1366/News-Comments-Breast-Cancer-Screening-v1

has also been extensively used for text clustering and topic modeling (Xu et al., 2003) (Shahnaz et al., 2006).

Online News Comment Analysis Automatic analysis of online news comments has been investigated in (Potthast et al., 2012), (Tsagkias et al., 2010). In (Zhou et al., 2010), different feature sets for sentiment analysis of news comments were compared. In (Chardon et al., 2013), the effect of using discourse structure for predicting news reactions was explored. In (Zhang et al., 2012), a supervised method for predicting emotions toward news such as sadness, surprise, and anger was proposed. Our paper is the first work toward stance classification of news comments which is particularly different from sentiment and emotion classification as stance is not necessary expressed by affective words and determining the polarity of the text is not sufficient since the system should detect favorability toward a specified target that may be different from the opinion target.

3 Dataset

Important results of health-related studies, reported in the scientific medical journals, are often popularized and broadcasted by media. Such media stories are often followed by online discussions in the social media. For our research, we chose to focus on a controversial study published in the British Medical Journal (BMJ) in February 2014, about breast cancer screening (Miller et al., 2014). Subsequently, a set of news articles that broadcasted or discussed about this study was selected and their corresponding comments were collected. There are two Yahoo news articles[2], three CNN[3] and three New York Times articles[4].

[2] 1.http://news.yahoo.com/mammograms-not-reduce-breast-cancer-deaths-study-finds-001906555.html
2.https://news.yahoo.com/why-recent-mammography-study-deeply-flawed-op-ed-170524117.html

[3] 1. http://www.cnn.com/2014/02/12/health/mammogram-screening-benefits/index.html
2.http://www.cnn.com/2014/02/19/opinion/welch-mammograms-canada/index.html
3.http://www.cnn.com/2014/03/18/opinion/sulik-spanier-mammograms/index.html

[4] 1.http://www.nytimes.com/2014/02/12/health/study-adds-new-doubts-about-value-of-mammograms.html,
2.http://www.nytimes.com/2014/02/15/opinion/why-i-never-got-a-mammogram.html,
3.http://well.blogs.nytimes.com/2014/02/17/a-fresh-case-for-breast-self-exams/

Comments were harvested from news websites or their corresponding social media. CNN commentsphere is provided by DISQUS[5]. Only root comments were kept and the rest (reply to the other comments) was discarded since they mostly contain user interactions and their opinion targets are not the study in which we are interested in for this research. A total number of 1063 posts were collected from all the sources and cleaned by removing HTML tags and links.

3.1 Annotation

Our annotation scheme consisted of two tasks: stance classification and argument tagging for each comment. For stance classification, we are interested in the overall position of the commenter toward the target medical research that is the BMJ article about breast cancer screening (Miller et al., 2014). Two possible positions toward this health-related study were considered:

- For/Agree/Support: those comments that are supporting the target study by arguing its pros or showing positive sentiments toward the target research or expressing their agreement. In other words, those commenters that react positively to the target research study.

- Against/Disagree/Opposition: those comments that are opposing the target study by arguing its cons or showing negative sentiments toward the target research or expressing their disagreement. In other words, those commenters that react negatively to the target research study.

In addition to the overall stance (for or against), we are interested in the strength of the position of commenters toward the target research. Thus, the annotators had five options to choose from: "Strongly For", "For", "Other", "Against", and "Strongly Against". Here, "Other" may correspond to neutral, ambiguous, or irrelevant comments. In opinion mining and sentiment analysis, it is essential to recognize what the opinion is about, which is called "opinion target". Irrelevant opinions may

[5] https://disqus.com

not be directly related to our target study. In this case study, we are interested in comments for which their opinion target is mammography/ breast cancer screening/the BMJ article. For instance, if the comment is about the reporter and the way she reports the research, it does not give us any information about the overall stance of the commenter toward the study. For some comments, it is impossible to judge the overall stance of commenters due to the lack of evidence/information about his/her position. This may also be due to a mixture of "for" and "against" arguments without any clear overall position. The annotator has labeled such comments as "Other", as they may be ambiguous or neutral.

We are not only interested in the overall position of commenter, but also in the reasons behind it. Commenters usually back up their stances with arguments. Our second annotation task was argument tagging in which the annotator identified which arguments have been used in a comment, from a predefined list of arguments. These tags are organized in a hierarchical tree-structured order, as some of them may be related. This structure is represented in figure 1. The annotators were instructed to choose leaf arguments (the most specific one) rather than more general ones, when possible. Commenters may use more than one argument to support their position. For this corpus, the annotators were asked to select at most two arguments based on the emphasis of the author on them. In other words, if the comment had more than two arguments, the ones with more emphasis were selected (because more than two arguments appeared in very few comments in our corpus). The predefined list of arguments was manually extracted and the annotators had chosen appropriate tags from this list, for each post.

Inter-annotator Agreement Our annotation consisted of two separate tasks. For each task, a different numbers of annotators have been used and the annotation was evaluated independently. Stance annotation was carried out by three annotators. To measure inter-annotator agreement, the average of weighted Kappa between each pair of annotators was calculated. As the labels have ordinal value and Fleiss' Kappa and Cohen's Kappa are mainly designed for categorical data, we did not use them to assess stance classification annotation. The major difference between weighted Kappa and Cohen's Kappa is that weighted Kappa considers the degree of disagreement.

	Weighted Kappa	Cohen's Kappa
Stance Classification (3-class)	0.62	-
Stance Classification (5-class)	0.54	-
Argument Tagging	-	0.56

Table 1: Inter-annotator agreement for argument tagging and stance classification

One annotator labelled the arguments for each post. However, to evaluate the quality of annotation, a subset of our corpus (220 comments) were selected and independently annotated by the second annotator. The annotations were compared without considering the hierarchical structure of the tags from figure 1. To measure inter-annotator agreement Cohen's Kappa was deployed. It is also possible to consider hierarchical structure of arguments and to calculate a weighted Kappa based on their distance in the tree.

Table 1 shows the inter-annotation agreement results for both tasks. The agreements are in the range of reported agreement in similar tasks and for similar data (Boltuzic and Šnajder, 2014) (Walker et al., 2012c). The values show the difficulty of the task, even for humans. Eventually, those comments for which at least two annotators agreed about the overall position (stance label) were kept and the rest, labeled as "Other" were discarded, as they may be truly ambiguous.

3.2 Corpus Analysis

As described earlier, our corpus has 1063 comments in total. After discarding those comments with stance label of "Other", 781 comments remained. Table 2 provides an overview of the stance labels in the corpus. The distribution of different argument tags over different stance labels is illustrated in table 3. Additionally, this table shows the number of occurrences of each argument in the corpus. As each comment has been annotated by two argument tags, the total is two times the number of comments. The number of "Other/None" labels is high because it was used as the second argument label for com-

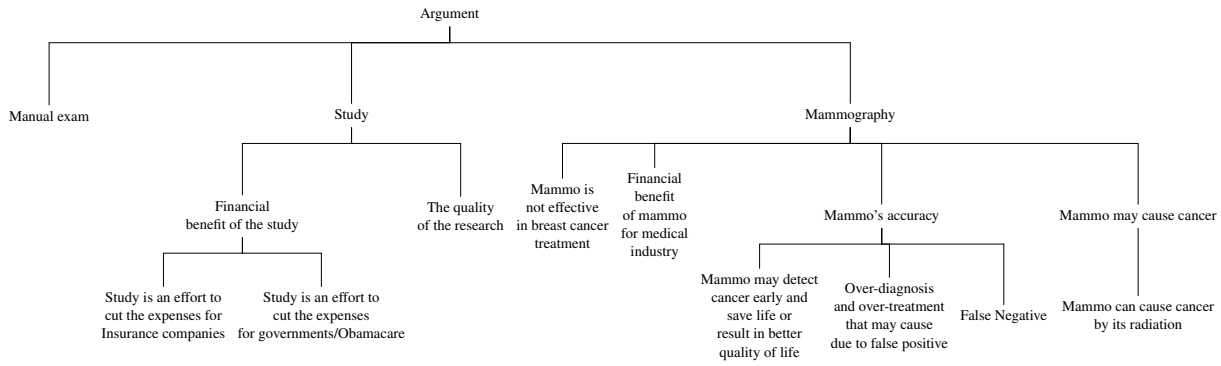

Figure 1: Hierarchical structure of arguments in our corpus

	Strongly For	For	Against	Strongly Against
Post	157	200	172	252

Table 2: Distribution of stance labels in the corpus

ments for which only one argument could be identified by the annotators. Because there are not sufficient instances in the corpus for some of the tags, and the data would be too imbalanced, we decided to remove tags that have less than five percent representatives in the corpus and replace them with the "Other/None" label.

4 Proposed Framework

In this paper, a framework for argument tagging is introduced. The main advantage of this framework is that labeled data is not required. In this approach, NMF is first applied on unlabeled data to extract topics. Subsequently, data are clustered based on these topics. Each post may belong to that topic cluster if its probability of generating from that topic is more than a certain threshold. Later, these clusters are labeled to match a predefined list of argument tags by an annotator. In summary, NMF can cluster comments based on their arguments and these clusters can be labeled by considering top keywords of each cluster topic.

To label each cluster, the top keywords of that topic and the list of arguments were given to the annotators. An annotator who is relatively familiar with comments can easily match topics with arguments, for any domain. The suggested framework for annotation is considerably less tedious and time consuming compared to annotating all posts one by one and leveraging them for training a supervised statistical learner. For our corpus, annotating all comments took 30 hour from for an annotator, while matching topics with argument tags took less than one hour. This illustrates the efficiency of the proposed framework.

In this framework, these extracted argument tags for each comment are subsequently leveraged for stance classification using an SVM classifier. Exploiting argument tags for predicting stance is beneficial, as an argument is often used to back up a single stance, either for or against.

5 Experiments and Results

In this section, first, the experimental setting is reviewed and the evaluation process and metrics are described. Subsequently, the results of applying our proposed framework on our corpus are presented for both argument tagging and stance classification.

5.1 Experimental Setup

After removing those arguments which did not have sufficient representatives, eight argument tags remained. We treated argument tagging as a multi-class multi-label classification problem. Each post can have one or more of those eight labels or none of them.

Each post was represented by using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme over its bag of words. Standard English stopwords were removed. Additionally, we removed corpus specific stopwords by discarding

Argument	Strongly For	For	Against	Strongly Against	Total
Argument about the study	0	1	1	1	3
The quality of the study	5	7	35	43	90
Financial benefit of the study	0	0	4	6	10
Study is an effort to cut the expenses for Insurance companies	0	2	22	26	50
Study is an effort to cut the expenses for governments/Obamacare	0	2	26	41	69
Argument about the mammography	2	1	0	0	3
Mammo is not effective in breast cancer treatment	5	9	1	2	17
Mammo may cause cancer	9	1	0	0	10
Mammo can cause cancer by its radiation	42	23	1	1	67
Mammo's accuracy	2	7	0	2	11
Over-diagnosis and over-treatment that may cause because of false positive	51	36	0	0	87
False Negative	13	17	1	0	31
Mammo may detect cancer early and save life or result in better quality of life	0	8	63	175	246
Financial benefit of mammo for medical industry	47	53	1	0	101
Argument about manual exam	20	29	10	9	68
Other/None	118	204	179	168	699
Total	314	400	344	504	1562

Table 3: Distribution of argument tags for different stance labels in our corpus

terms that have been appeared in more than twenty percent of the documents.

For evaluation, separate test and training data were deployed. Data was randomly divided into test and training sets. Seventy percent of the data was used for training and the rest was used for testing. As mentioned earlier, for our proposed framework, the labels of training are not leveraged and topic models are applied on unlabeled training data. Like similar researches in text classification, precision, recall and f1-score are used as evaluation metrics.

5.2 Argumentation Mining Results

In this section, the results of applying our proposed framework are described and compared to a supervised classifier that uses the same features (TF-IDF). As a supervised classifier, a linear multi-label Support Vector Machine (SVM) is employed using the one-versus-all training scheme. Additionally, in our framework instead of NMF, LDA was used for topic modeling and the results are compared between the two approaches.

The number of topics for our topic models is set to the number of argument tags. As mentioned earlier, after removing those tags with insufficient data, eight arguments remained. These topics, represented by their top keywords, were given to two annotators and we asked them to match them with the list of arguments. Another advantage of the NMF topics is that in this case, both annotators were agreed on all labels. The topics extracted by LDA were difficult for annotators to label, as they were vague. The annotators agreed on fifty percent of labels (4 out of 8 labels). To be able to make a decision in the cases of disagreement, we asked a third annotator to choose one of the suggested labels by two other annotators. Table 4 shows the eight argument tags and their matched NMF and LDA topics, as represented by their top keywords.

Argument	NMF Topic	LDA Topic
1) The quality of the study	study, death, mammography, group, rate, survival, canadian, quality, woman, data, result, question, poor, medical, used, better, trial	insurance, want, company, age, test, early, treatment, screen, write, doctor, thing, benefit, need, unnecessary, group, family, earlier, stage
2) Study is an effort to cut the expenses for insurance companies	insurance, company, pay, cover, sure, way, funded, maybe, wait, ploy, wonder, procedure, benefit, provide, expensive, worth, make, money	saved, insurance, health, care, screening, save, company, money, healthcare, doctor, mammography, exam, self, like, responsible, expensive
3) Study is an effort to cut the expenses for governments/Obamacare	obamacare, drop, test, past, paid, cut, obama, change, socialized, waste, ordered, future, routine, bad supposed, trying, notice, lady, cost	think, test, early, better, obamacare, money, self, treatment, screening, insurance, exam, article, medical, detect, make, told, decision, yearly
4) Mammo can cause cancer by its radiation	radiation, lumpectomy, expose, need, colonoscopy, surgery, chemo, cause, radiologist, machine, treatment, exposure, safe, thermography	know, radiation, mammography, cut, data, radiologist, tumor, need, surgery, medical, early, maybe, really, time, getting, exam, waited, way
5) Over-diagnosis and over-treatment that may cause due to false positive	medical, false, psa, risk, needle, biopsy, screening, prostate, positive, research, surgery, factor, best, painful, over, diagnosis, needed, died	treatment, think, radiation, stage, like, make, yearly, time, article, came, test, doctor, biopsy, self, mother, screening, psa, survivor, lump
6) Mammo may detect cancer early and save life or result in better quality of life	saved, stage, diagnosed, routine, early, today, discovered, mother, believe, alive, friend, annual, detect, late, aggressive, regular	stage, radiation, saved, doctor, early, later, screening, result, want, stop, treatment, like, invasive, happy, routine, mammography, patient, diagnostic
7) Financial benefit of mammo for medical industry	money, care, healthcare, medicine, people, cost, screening, preventive, responsible, administration, way, let, control, doctor expensive, industry	medicine, doctor, treatment, radiation, death, early, catching, money, save, needle, detection, test, making, saved, u, canada, mammography, form
8) Argument about manual exam	exam, self, lump, tumor, physical, manual, regular, examination, time, malignant, trained, nurse, rely, survivor, fast, yes, detecting change	know, people, hope, health, let, need, want, tumor, pay, radiation, like, death, dci, test, alive, exam, age, look, saved, doctor, evidence, say, human

Table 4: Extracted topic by NMF and LDA models represented by their top keywords

	Precision	Recall	F1-score
Linear-SVM	0.76	0.33	0.43
Cluster-LDA	0.26	0.32	0.28
Cluster-NMF	0.58	0.53	**0.49**

Table 5: Results of argument tagging on our corpus

	Precision	Recall	F1-score
Baseline	0.16	0.40	0.23
TF-IDF	0.43	0.45	0.37
TF-IDF+Args	0.48	0.48	**0.47**

Table 6: Results of stance classification in the case of 4-classes (the strength and the overall stance)

Table 5 presents the precision, recall and f1-score of the argument tagging task on our corpus. Our model based on NMF outperforms the other two approaches significantly in term of f1-score and recall, while it is considerably more efficient in terms of the required annotation.

5.3 Stance Classification Results

For stance classification, the predicted argument tags from the previous section were leveraged for stance classification. Our proposed stance classifier deploys the same set of TF-IDF features; in addition, it uses the predicted argument tags as features and as a classification method, linear SVM is employed. These methods are compared with two other classifiers: a linear SVM with TF-IDF as features, and a simple majority class classifier as a baseline. The results are shown in two settings.

Table 6 presents the results of predicting both the stance and its strength (4-class), while table 7 shows the result of stance classification (for or against). Comments with the label of "Other" have been already removed from data. In both settings, the performance is improved when adding the predicted arguments as features.

6 User Generated Content Visualization

In this section, one of the applications of automatic analysis of news comments is illustrated. Following the extraction of arguments from news comments, they can be visualized. In figure 2, the distribution of main arguments in the corpus based on the human annotation are represented, while in figure 3 the distribution based on the automatically-predicted arguments is demonstrated. The figures visualize the relative importance of the arguments. Such visualizations could be really useful to decision makers, even if the arguments were automatically predicted, therefore not all the predictions are correct, because their relative importance was correctly detected. Most importantly, the predictions can be obtained for any domain by using our method, without

	Precision	Recall	F1-score
Baseline	0.32	0.56	0.41
TF-IDF	0.79	0.76	0.74
TF-IDF+Args	0.77	0.77	**0.77**

Table 7: Results of stance classification in the case of 2-classes

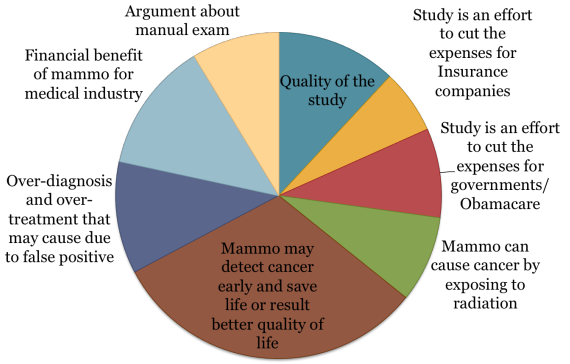

Figure 2: The summary of arguments based on annotated data

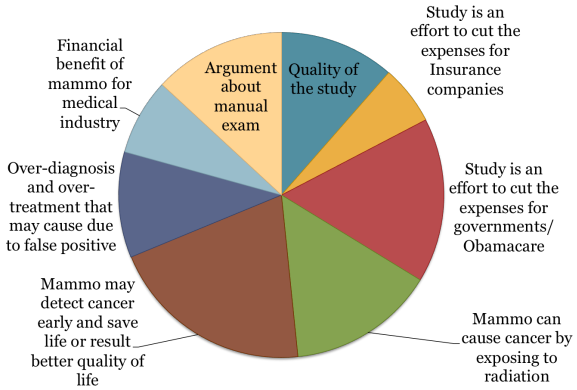

Figure 3: The summary of arguments based on predicted data

the need to label large amounts of data.

7 Discussion

In this section, we further investigate and analyze the results presented earlier. In the previous section, it was shown that using NMF for clustering comments based on their arguments is significantly better than employing LDA. This can be observed in the extracted top keywords of the topics. NMF topics can be matched to the arguments considerably more easily. This is also supported by the evaluation results, as clustering based on NMF has significantly better precision, recall, and f1-score than clustering using LDA. We speculate that the reason for this is the shortness of the comments, since LDA normally works better for longer texts. The other reason may be the fact that all of these data are about the same general topic, breast cancer screening, and LDA cannot distinguish between subtopics (different arguments).

Table 6 demonstrates that stance prediction is significantly improved by leveraging the predicted argument tags. The reason for this can be simply explained by referring to table 3. This table shows that most of the arguments have been leveraged mainly to back up a single stance. Hence, by predicting the correct argument, the stance can be guessed with high probability. The correlation between stance labels and argument tags has been also observed in (Boltuzic and Šnajder, 2014), but they have exploited manually-annotated stance labels for argument classification.

To explore in more details the results of our proposed framework, precision, recall and f1-score for each class (argument tag) is illustrated in table 8. Better precision is achieved for argument classes that are more explicitly expressed and similar sentences are used to convey them. The argument "Mammo may detect cancer early and save life or result in better quality of life" (class 6) has the best precision, as it is mostly expressed by sentences like "Mammography saved my/my mother/ my friend life". On the contrary, our method has better recall for those arguments referred more implicitly in the corpus. For instance, the argument class "Study is an effort to cut the expenses for governments/Obamacare" (class 4) has low precision and high recall, due to several posts such as " Step in the direction of limited health care. You know, hope and change." that implicitly express this argument. Another reason for low precision of some classes, such as "Argument about manual exam" (class 8), is that the corpus is imbalanced and they have less representative data compared to others.

Class	Cluster-NMF		
	Precision	Recall	F1-score
1	0.34	0.61	0.44
2	0.56	0.83	0.67
3	0.57	0.24	0.33
4	0.33	0.68	0.44
5	0.40	0.50	0.44
6	0.91	0.38	0.54
7	0.44	0.65	0.52
8	0.39	0.71	0.51

Table 8: The summary of the performance of proposed framework for each argument (the class numbers match argument tag numbers in table 4)

8 Conclusion and Future Work

Stance classification and argumentation mining were recently introduced as important tasks in opinion mining. There has been a growing interest in these fields, as they can be advantageous particularly for decision making. In this paper, a novel framework for argument tagging was proposed. In our approach, news comments were clustered based on their topics extracted by NMF. These clusters were subsequently labeled by considering the top keywords of each cluster.

The main advantage of the proposed framework is its significant efficiency in annotation. Most of the previous works required a large set of annotated data for training supervised classifiers, and the annotation process is tedious and time-consuming, while in our approach there is no need for labeled training data for the argument detection task. The annotation needed for the argument detection task is minimal: we only need to map the automatically-detected topics to the arguments. This mapping can be easily done for new subjects. Considering the huge amount of news comments that are generated every day for various subjects, this advantage is significant.

Several lines of research can be investigated in the future. First, we plan to apply our framework on available datasets for argument tagging and stance classification of ideological debates. to study its performance in other domains. Furthermore, we intend to concentrate more on the hierarchical structure of the argument tags, by exploiting hierarchical topic modeling to extract arguments with different levels of abstractness. Another area that can be explored is automatic extraction of the set of argument tags, in a similar way to the automatic aspect extraction of product reviews.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada under the CREATE program, and of the Polish National Scientific Centre NCN grant UMO-2013/09/B/ST6/01549. We thank Kenton White for motivating us to employ NMF for topic modeling. We thank our annotators Raluca Tanasescu and Nasren Musa Elsageyer.

References

Camille Albert, Leila Amgoud, Florence Dupin de Saint-Cyr, Patrick Saint-Dizier, and Charlotte Costedoat. 2011. Introducing argumention in opinion analysis: Language and reasoning challenges. *Sentiment Analysis where AI meets Psychology (SAAIP)*, page 28.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 1–9. Association for Computational Linguistics.

Frank R Baumgartner, Suzanna L De Boef, and Amber E Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. Cambridge University Press.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Filip Boltuzic and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Baptiste Chardon, Farah Benamara, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013. Measuring the effect of discourse structure on sentiment analysis. In *Computational Linguistics and Intelligent Text Processing*, pages 25–37. Springer.

Alexander Conrad, Janyce Wiebe, et al. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88. Association for Computational Linguistics.

Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48.

Kazi Saidul Hasan and Vincent Ng. 2013. Frame semantics for stance classification. *CoNLL-2013*, 124.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. *EMNLP 2014*.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, Andrew Ravenscroft, and David Bourget. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. *ACL 2014*, page 79.

Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 109–116. Association for Computational Linguistics.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM.

Anthony B Miller, Claus Wall, Cornelia J Baines, Ping Sun, Teresa To, Steven A Narod, et al. 2014. Twenty five year follow-up for breast cancer incidence and mortality of the canadian national breast screening study: randomised screening trial. *Bmj*, 348.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Martin Potthast, Benno Stein, Fabian Loose, and Steffen Becker. 2012. Information retrieval in the commentsphere. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):68.

Farial Shahnaz, Michael W Berry, V Paul Pauca, and Robert J Plemmons. 2006. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

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

Ivan Titov and Ryan McDonald. 2008a. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

Ivan Titov and Ryan T McDonald. 2008b. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, pages 308–316. Citeseer.

Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. 2010. News comments: Exploring, modeling, and online prediction. *Advances in Information Retrieval*, pages 191–203.

Marilyn A Walker, Pranav Anand, Rob Abbott, Jean E Fox Tree, Craig Martell, and Joseph King. 2012a. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012b. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596. Association for Computational Linguistics.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012c. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60. Association for Computational Linguistics.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM.

Ying Zhang, Yi Fang, Xiaojun Quan, Lin Dai, Luo Si, and Xiaojie Yuan. 2012. Emotion tagging for comments of online news by meta classification with heterogeneous information sources. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1059–1060. ACM.

Jie Zhou, Chen Lin, and Bi-cheng Li. 2010. Research of sentiment classification for net news comments by machine learning. *Journal of Computer Applications*, 30(4):1011–1014.