Computational Models of Argument
S. Modgil et al. (Eds.)

doi:10.3233/978-1-61499-906-5-161

Classifying Types of Ethos Support and Attack

Rory DUTHIE [a,1], and Katarzyna BUDZYNSKA [b,a]
[a] *Centre for Argument Technology, University of Dundee, UK*
[b] *Centre for Argument Technology, Institute of Philosophy and Sociology, Polish Academy of Sciences, Poland*

Abstract. Endorsing the character of allies and destroying credibility of opponents is a powerful tactic for persuading others, impacting how we see politicians and how we vote in elections, for example. Our previous work demonstrated that ethos supports and attacks use different language, we hypothesise that further distinctions should be made in order to better understand and implement ethotic strategies which people use in real-life communication. In this paper, we use the Aristotelian concept of elements of ethos: practical wisdom, moral virtue and goodwill, to determine specific grounds on which speakers can be endorsed and criticised. We propose a classification of types of ethos supports and attacks which is empirically derived from our corpus. The manual classification obtains a reliable Cohen's kappa $\kappa = 0.52$ and weighted $\kappa = 0.7$. Finally, we develop a pipeline to classify ethos supports and attacks into their types depending on whether endorsement or criticism is grounded in wisdom, virtue or goodwill. The automatic classification obtains a solid improvement of macro-averaged F1-score over the baseline of 10%, 25%, 9% for one vs all classification, and 16%, 18%, 10% for pairwise classification.

Keywords. Corpus Analysis, Ethos-Logos, Ethos Mining, Elements of Ethos, Wisdom-Virtue-Goodwill

1. Introduction

Mining arguments (*cf.* [5]) has become a rapidly growing area of AI which builds upon methods and techniques from sentiment analysis and opinion mining (*cf.* [15]) and incorporates tasks such as argument scheme classification [8,13]. With both academic research labs and engagement from commercial R&D such as the IBM Watson Debater project team [2], results have started to make significant headway against the challenge of content-based understanding of arguments, i.e. *logos*. In this work, we look into *ethos*, the character of the speaker [1, 1356a]. We build upon our previous work of annotation and automation of ethos supports and attacks from UK parliamentary debates [7,6], investigating on what grounds politicians endorse and criticise each other. The examples below show three common ethotic strategies: in Example 1 Mr. Moore is supporting the experience and knowledge of an entity (Miss Widdecombe); in Example 2 Mr. Jenkin is endorsing the Government for having courage; and in Example 3 Mr. Moore is referring

[1]Corresponding Author: School of Science and Engineering, University of Dundee, Nethergate, Dundee, DD1 4HN, United Kingdom; E-mail: rwduthie@dundee.ac.uk

to Mr. Meyor's good deeds in respect to an audience (his constituents). These strategies correspond to three elements of ethos studied in rhetoric: practical wisdom referring to having knowledge; moral virtue when knowledge is revealed (i.e. honesty); and goodwill when the knowledge is shared (i.e. giving the best advice to others) [1, 1378a].

Example 1 Mr. John Moore said, *I bow to my hon. Friend's distinguished past and detailed knowledge of these matters.*

Example 2 Mr. Patrick Jenkin said, *I believe that the Government were right to have the courage to bring forward the necessary measures to bring public expenditure under control.*

Example 3 Mr. John Moore said, *My hon. Friend is assiduously pursuing his constituents' interests.*

In this paper, we first re-annotated the corpus developed in our previous work [7,6] which contains ethos supports (+ESE a support of another entity's ethos, i.e. politician or party) and ethos attacks (-ESE an attack on another entity's ethos) (see Section 2.2). We applied the distinction of wisdom, virtue and goodwill, following a modified model (for annotation purposes) of ethos proposed in rhetoric (Section 2.3.1). The final corpus contains six types of ethotic structures for support (Argument from Practical Wisdom; Moral Virtue; and Goodwill) and attack (Conflict from Practical Wisdom; Moral Virtue; and Goodwill). Then, we developed a pipeline for automatic classification of these six types, building upon the ethos mining technology introduced in [7,6] (Section 3.2). Positive and negative ethotic sentiment expressions, +/- ESEs, are used as an input to the pipeline which are broken into features. The pipeline is then evaluated using One vs All classification, to determine the distinction between types of ethos, and pairwise classification, to determine the overall outcome when a data point is classified in more than one category (Section 3.3). Specifically this paper contributes: (*a*) the first freely available corpus of manually classified types of ethos support and attack; (*b*) freely available guidelines for applying Wisdom, Virtue and Goodwill tags; (*c*) a pipeline for automatic classification of ethos support and attack types (wisdom, virtue and goodwill).

2. Manual Classification

In this section we outline the related work for the manual classification of ethos types, then our previous work and finally our annotation process.

2.1. Related work

Ethos has been incorporated into philosophy and informal logic and studied as ethotic argument (*cf.* [3,19,11]) and interpersonal argumentation [4]. One of the most popular ethotic structures are specified for support as Arguments from Position to Know in Figure 1(a) (and its versions Expert Opinion; Witness Testimony; and Deontic Authority); and for attack as forms of Ad Hominem (AH) argument with its simplest version of Generic AH in Figure 2(a). This is extended to allow the ability to infer a propositional content from any speech act performed by an authority (Figure 1(b)) and to allow the representation of an attack in the structure of AH argument (as conflict, see Figure 2(b)).

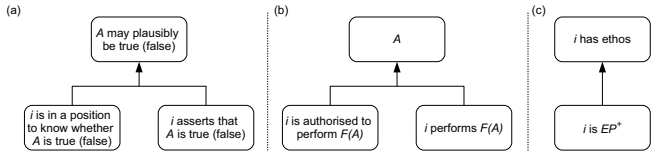

Figure 1. Ethotic structures with support relation in: (a) Argument from Position to Know [19]; (b) a model with a speech act $F(A)$ [4]; and (c) a model with a positive ethotic property, EP^{+} [7,6].

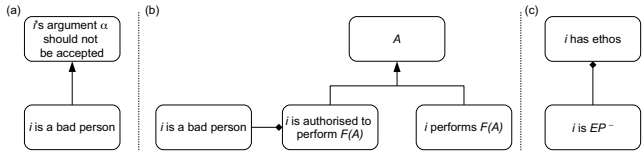

Figure 2. Ethotic structures with attack relation in: (a) Generic Ad Hominem, AH [19]; (b) AH model with a speech act $F(A)$ and attack relation (graphically represented differently than inference relation) [4]; and (c) a model with a negative ethotic property, EP^{-} [7,6].

In our approach ethos plays a different role in the structure: in Figures 1(a) and 1(b) ethos supports a proposition, while in 1(c) ethos is supported (ethos can be used to support what a speaker is saying or general ethos). Similarly, in Figures 2(a) and 2(b) ethos attacks an argument or proposition, while in 2(c) ethos is attacked.

In [9] 200 AH instances, from Change My View on Reddit, were annotated for types (abusive, tu quoque, circumstantial, bias and guilt by association) using mechanical turk. This annotation achieved a low percentage agreement (score reported as low) where the authors highlighted that the AH arguments within reddit did not fall under distinct classes in the theory, due to multifaceted cases. In [14], annotation was conducted on reputation defence strategies (denial, excuse, justification, concession and no strategy) in the Canadian Hansard. Question, answer pairs were extracted and given to multiple annotators where agreements on pairs were included in the dataset (493 pairs overall).

2.2. Ethos supports and attacks

The annotation of types of ethos supports and attacks builds upon that of ethos which was undertaken in our previous works [7,6]. Hansard, the UK parliamentary debate record, was used to obtain transcripts for annotation with sixty used in [7] and ninety in [6].

Annotation scheme. Four tags (see [7] and [6] for occurrences) were used for annotation: source, the speaker of an ethotic statement; target, the referent of an ethotic statement; ethos support, +ESE where: (*a*) the statements refers to a person or group explicitly; and (*b*) it supports their credibility or looks to put them in a positive frame; and (*c*) it is not a reference to one's own credibility; ethos attack -ESE: when the opposite values hold true. All other statements are non-Ethotic sentiment expressions (n-ESEs). In [6], ethotic statements must contain some form of linguistic surface as an indicator to ethos.

Corpora. The first corpus of annotated ethotic support and attack relations was EtHan_Thatcher_3 (`http://arg.tech/Ethan_Thatcher`), containing 60 transcripts, split into training and test data. This was extended to 90 transcripts in the

Tags	Ethos Supports	Ethos Attacks	Total	Word Count
Wisdom	48 (29%)	190 (40%)	238 (37%)	6,954
Virtue	99 (59%)	194 (41%)	293 (46%)	7,611
Goodwill	20 (12%)	87 (19%)	107 (17%)	3,685
Total	167 (100%)	471 (100%)	638 (100%)	18,250

Table 1. Occurrences of tags with the respective word counts for each segment in EthosWVG_Hansard.

Ethos_Hansard corpus (`http://arg.tech/Ethos_Hansard`) where the training data was extended. Both corpora were annotated using the OVA+ annotation tool [10] which incorporates the Argument Interchange Format (AIF) [17] for argument representation and were subsequently stored in the AIFdb [12] (`http://aifdb.org`).

Evaluation. In the case of both corpora 10% was annotated by a second annotator to validate the guidelines through inter-annotator agreement (IAA) using Cohen's kappa. In EtHan_Thatcher_3 $\kappa = 0.67$ for ESEs and n-ESEs, $\kappa = 0.95$ for +/- ESE combined, $\kappa = 1$ for source and $\kappa = 0.84$ for the target. In Ethos_Hansard $\kappa = 0.67$, $\kappa = 1$ for +ESE and -ESE combined, $\kappa = 1$ for source and $\kappa = 0.93$ for the target.

2.3. Types of ethos supports and attacks: wisdom, virtue and goodwill

The three ethos types (wisdom, virtue and goodwill) are annotated using OVA+ and annotation stored in AIFdb where ethos types use the base annotation from [6].

2.3.1. Classification of elements of ethos

Annotation scheme. The annotation of the ethos types are broken into two categories knowing information (knowledge) or knowing the right actions (actions). Moral Virtue and Goodwill are further distinguished, when the categories apply in general this is virtue when applied to an audience this is goodwill. The tags are applied in the guidelines (see `http://arg.tech/WVGGuideNew` for full guide) as follows:

Practical Wisdom. Argument From Practical Wisdom is annotated when an entity: (*a*) knows the right information; or (*b*) knows the right action. Conflict From Practical Wisdom is annotated when an entity: (*a*) does not know the right information; or (*b*) does not know the right action.

Moral Virtue. Argument From Moral Virtue is annotated when an entity: (*a*) knows and reveals the right information in general; or (*b*) is honest in general; or (*c*) performs the right action when they know it; or (*d*) does the right action in general. Conflict From Moral Virtue is annotated when an entity: (*a*) knows information but does not reveal it in general; or (*b*) lies in general; or (*c*) performs an action when they know it is wrong; or (*d*) does the wrong action in general.

Goodwill. Argument From Goodwill is annotated when an entity: (*a*) knows and shares information with the audience; or (*b*) is honest with the audience; or (*c*) performs the right action for others aligning with their values giving sound advice; or (*d*) does not do wrong to others. Conflict From Goodwill is annotated when an entity: (*a*) does not share information with the audience; or (*b*) misleads the audience; or (*c*) does not do what they know is right for the audience; or (*d*) does the wrong things for an other or audience.

Corpus. Following the annotation of wisdom, virtue and goodwill the publicly available EthosWVG_Hansard corpus (`http://arg.tech/EthosWVG`) was created. The corpus builds upon the tags in the Ethos_Hansard corpus to create the only available corpus of ethos types, containing 638 segments and 18,250 words (see Table 1 for details).

Evaluation. To evaluate the guidelines, a 12% subset was annotated by a second annotator giving $\kappa = 0.52$, a percentage agreement of 66% and weighted $\kappa = 0.70$. Combining ethos types into polynomial classes gives an idea of the disagreements between the annotators for wisdom and virtue against goodwill $\kappa = 0.81$. While virtue and goodwill against wisdom $\kappa = 0.61$. This shows that there is a clear distinction between virtue and goodwill although the difference between wisdom and virtue could be improved.

Example 4 *I understand the right hon. Gentleman to be something of an expert in doggerel and verse.*

Further error analysis showed that the the distinction between wisdom and virtue can be difficult.In Example 4 the entity is referred to as an expert, pertaining to wisdom, however, doggerel and verse is considered as clumsy and irregular, an attack from virtue.

3. Automatic Classification

In this section we outline the related work for the automatic classification of ethos types, our previous work and finally our new classification pipeline and evaluation.

3.1. Related work

Although classifying argument schemes substantially differs from ethos types the techniques used are similar. In [8], five argumentation schemes are identified (example, cause to effect, practical reasoning, consequence and verbal classification) and classified using a One vs All classification. Data from the Aracaria database was used to train a decision tree (accuracy ranging from 63.2% to 90.8%). In [13] four schemes were classified (analogy, cause to effect, practical reasoning and verbal classification) using One vs All classification and arguments from AIFdb where a Naive Bayes classifier F1-scores averaging between 0.65 to 0.75. In [9], a convolutional neural network was used to extract cases of AH from reddit Change My View threads. 7,242 instances of AH were extracted for training and testing purposes and gave an accuracy of 81%. In [14](see Section 2.1 for annotation details), reputation strategies were classified using both mutliclass classification and pairwise classification with an SVM giving an F1 score of 0.57.

3.2. Ethos supports and attacks

The automatic extraction of ethos supports and attacks was undertaken in [7] (see Section 3.2.1) and [6] (see Section 3.2.2) of which the classification of types builds upon.

3.2.1. Domain Rules

In [7], ethos supports and attacks were classified using a pipeline of natural language processing (NLP) techniques (including POS tags, anaphora resolution and reported speech removal to extract ESEs and sentiment classification for +/- ESEs). The ESE/n-ESE classification gave an F1-score of 0.70 and the +/- ESE an F1-score of 0.78.

3.2.2. Deep learning

In [6], a pipeline of NLP techniques classified ethos supports and attacks. Raw text is passed to: a Deep Modular Recurrent Neural Network (DMRNN); the Stanford POS tagger; the Stanford Universal Dependency (UD) tagger [18]; a sentiment classifier; and anaphora resolution. POS tags are then passed to the DMRNN. The UD tags are passed to the DMRNN and to entity extraction (EXT) where the output is passed to the DMRNN. A sentiment classifier determines polarity (F1-score 0.84) and polarity tags are passed to the DMRNN. The DMRNN (described in [6]) classifies ESEs/n-ESEs (macro-F1 0.83) and finally sentiment and anaphora resolution give the source, target and +/- ESEs.

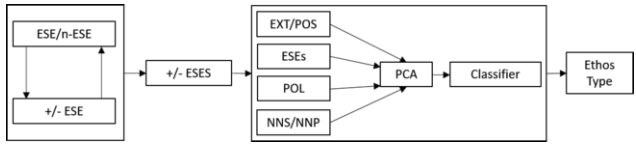

Figure 3. Pipeline for classifying types of ethos support and attacks containing a combination of entity relations and POS tags, ESEs, POL for ESEs and the presence of NNS and NNP tags. These are passed to a PCA module to reduce the dimensionality of the data for classification, which ultimately gives an ethos type.

3.3. Types of ethos supports and attacks: wisdom, virtue and goodwill

Implementation. To classify ethos support and attack types, an extension was made to the existing pipeline for +/- ESE classification (see Figure 3). +/- ESEs are broken into separate features: entity relations (EXT) from [6] are combined with POS tags for a new EXT/POS module; the raw ESE text; ESE polarity; and a plural and proper nouns (NNS/NNP) presence module. These were passed to a principal component analysis (PCA) module after-which a classifier determines the ethos support and attack type.

EXT / POS. Entities which are not relevant to ethos and the words related to them are removed (e.g. "Will the hon. Member"). Manual domain rules are executed upon UD tags (see [6] for explanation) and POS tags are matched to the entities and words.

ESEs. The ESE text has stop words removed and is transformed into unigrams, bigrams and trigrams as features for classification.

POL. All data points are split into positive and negative due to the varying language that can be used to support and attack.

NNS / NNP. Binary labels are created where an ESE contains or does not contain a plural or proper noun. The intuition being that goodwill instances, will mainly feature an additional entity (e.g. "constintuents") which is not captured using standard NER tools.

PCA. Scikit-learn's [16] built in PCA is used to make the data linearly separable where fifteen component features were used as points for the reduction to a 2D space.

Classifier. Four classifiers were considered for classification: Linear Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB) and a Decision Tree (DT). Scikit-learn was used for the classifiers all performing both classifications.

	Wisdom				Virtue				Goodwill			
	P	R	F1	m-F1	P	R	F1	m-F1	P	R	F1	m-F1
Baseline	0.34	0.35	0.35	0.51	0.44	0.45	0.44	0.48	0.15	0.16	0.15	0.71
SVM	0.42	0.68	**0.52**	0.54	0.58	0.52	**0.55**	**0.60**	0.27	0.50	0.35	0.72
LR	0.43	0.61	0.50	**0.56**	0.55	0.54	**0.55**	0.59	0.29	0.55	**0.38**	0.73
NB	0.39	0.49	0.43	0.53	0.51	0.57	0.54	0.55	0.25	0.36	0.30	0.73
DT	0.43	0.49	0.43	0.53	0.50	0.55	0.53	0.54	0.31	0.31	0.31	**0.77**

Table 2. One vs All classification for Wisdom, Virtue and Goodwill. Precision, recall, F1-score and macro-averaged F1-score are reported where the F1-score relates to the type in question and macro-averaged F1-score the combined classification. A baseline classifying on the class distributions is compared against machine learning classifiers using a 10-fold cross validation. Bolding denotes the classifiers with the highest scores.

	Wisdom / Virtue			Wisdom / Goodwill			Virtue / Goodwill		
	P	R	F1	P	R	F1	P	R	F1
Baseline	0.49	0.48	0.49	0.56	0.56	0.56	0.62	0.60	0.61
SVM	0.54	0.55	0.55	0.65	0.67	**0.66**	0.67	0.67	**0.67**
LR	0.57	0.58	**0.57**	0.67	0.66	**0.66**	0.71	0.65	**0.67**
NB	0.54	0.54	0.54	0.61	0.62	0.62	0.68	0.64	0.65
DT	0.56	0.56	0.56	0.64	0.65	0.65	0.66	0.67	0.66

Table 3. Pairwise classification results for Wisdom / Virtue, Wisdom / Goodwill and Virtue / Goodwill. Macro-averaged precision, recall and F1-score are reported for a 10-fold cross validation. A baseline classifying on the class distributions is compared against machine learning classifiers. Bolding denotes the highest F1-scores.

Evaluation. Results are reported for One vs All classification and pairwise classification following the procedure in [13] and [8]. In Table 2 (One vs All classification), SVM, LR, NB and DT classifiers are compared against a baseline classifying on the class distributions. A 10-fold cross validation is used for testing with the SVM and LR performing the best but all classifiers above the baseline. The highest F1-score of 0.55 was on the virtue class by the SVM (25% over the baseline). The highest macro-averaged F1-score of 0.77 was on the goodwill class by the DT although the F1-score was only 0.31.

In Table 3 (pairwise classification), SVM, LR, NB and DT classifiers are compared with a baseline classifying on the training set distributions. A 10-fold cross validation was used to test the classifiers all performing above the baseline and LR performing best overall. The results show a distinction between goodwill against wisdom and virtue (F1 = 0.67), but, the distinction between wisdom and virtue (F1 = 0.57) is less clear.

Errors pertain to the coupled nature of ethos types. In goodwill classification, the language is similar to virtue with entity mentions the difference, our NNS/NNP module and NER tools are not precise enough, so a new solution is needed. Similarly a domain context is needed to realise the difference between wisdom and virtue (see Example 4).

4. Conclusion

Research on the annotation and extraction of ethos has increased, but, there is still a lack of understanding of ethotic statements. In this paper we presented the first corpus and manual classification guidelines and the first automatic classification of wisdom, virtue and goodwill. We have robustly evaluated our annotation process ($\kappa = 0.52$) and our automatic classification pipeline shows promising results (F1 scores ranging from 0.53 to

0.77) but ethos support and attack types pose new questions. Can ethos types aid in ethos mining (investigated for argument schemes [13])? What situations are ethos types used in? While future work can address this, our paper has investigated a step in the direction to understanding ethos supports or attacks through wisdom, virtue and goodwill.

Acknowledgements

This work was supported in part by EPSRC in the UK under grant EP/M506497/1 and in part by the Polish National Science Centre under grant 2015/18/M/HS1/00620.

References

[1] Aristotle. *On Rhetoric (G. A. Kennedy, Trans.)*. New York: Oxford University Press., 1991.
[2] Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, 2017.
[3] A. Brinton. Ethotic argument. *History of Philosophy Quarterly*, 3:245–257, 1986.
[4] Katarzyna Budzynska. Argument analysis: Components of interpersonal argumentation. In *Proc. of COMMA*, pages 135–146, 2010.
[5] Katarzyna Budzynska and Serena Villata. Argument mining. In *The IEEE Intelligent Informatics Bulletin*, volume 17, pages 1–7, 2016.
[6] Rory Duthie and Katarzyna Budzynska. A Deep Modular RNN Approach for Ethos Mining. *To Appear in Proc. of IJCAI*, 2018.
[7] Rory Duthie, Katarzyna Budzynska, and Chris Reed. Mining Ethos in Political Debate. In *Proc. of COMMA*, pages 299–310. IOS Press, Berlin, 2016.
[8] Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In *Proceedings of the The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2011)*, pages 987–996, 2011.
[9] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. *CoRR*, abs/1802.06613, 2018.
[10] M. Janier, J. Lawrence, and C. Reed. OVA+: An argument analysis interface. In *COMMA*, 2014.
[11] M. Koszowy and D. Walton. Epistemic and deontic authority in the argumentum ad verecundiam. *Pragmatics and Society*, 2018.
[12] J. Lawrence, M. Janier, and C. Reed. Working with open argument corpora. In *ECA*, 2015.
[13] J. Lawrence and C.A. Reed. Argument mining using argumentation scheme structures. In P. Baroni, M. Stede, and T. Gordon, editors, *Proc. of COMMA*, Berlin, 2016. IOS Press.
[14] Nona Naderi and Graeme Hirst. Recognizing reputation defence strategies in critical political exchanges. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 527–535, 2017.
[15] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2008.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
[17] I. Rahwan, F. Zablith, and C. Reed. Laying the foundations for a world wide argument web. *Artificial Intelligence*, 171:897–921, 2007.
[18] Sebastian Schuster and Christopher D. Manning. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. *In LREC 2016*, pages 2371–2378, 2016.
[19] Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation schemes*. Cambridge University Press, 2008.