

# The Knowledge of Good and Evil: Multilingual Ideology Classification with PARAFAC2 and Machine Learning

PETER CHEW

*Hoffmantown Church, Albuquerque, NM, USA*

PHILIP KEGELMEYER

BRETT BADER

*Sandia National Laboratories, Albuquerque, NM, USA*

AHMED ABDELALI

*New Mexico State University, Las Cruces, NM, USA*

## ABSTRACT

*We explore the problem of automated classification of multilingual documents by ideology; that is, whether machine learning techniques can be applied to draw conclusions about the types of belief an author holds based just on text he/she produces. For training data, we use documents in a variety of languages. Some are collected from the world wide web by a “spider”, while others are hand-picked “ideological” documents, such as the writings of Lenin and Hitler. The documents are projected into a single cross-language semantic space defined by the application of PARAFAC2 to a multi-parallel aligned corpus. We find that standard machine learning techniques can be used to distinguish “ideological” documents from general web content with a high rate of accuracy (higher than has been reported for sentiment analysis problems), but the accuracy is slightly lower in distinguishing between ideologies. We conclude by discussing the factors that may contribute to our findings.*

## 1. INTRODUCTION

It is almost a truism by now to say that the amount of information available today as on-line text is huge. As the volume of available information has grown, so has interest in methods for automated