Content-Based Conflict of Interest Detection on Wikipedia

Udochukwu Orizu, Yulan He
School of Engineering and Applied Science, Aston University, UK
{orizuus,y.he9}@aston.ac.uk

Abstract

Wikipedia is one of the most visited websites in the world. On Wikipedia, Conflict-of-Interest (CoI) editing happens when an editor uses Wikipedia to advance their interests or relationships. This includes paid editing done by organisations for public relations purposes, etc. CoI detection is highly subjective and though closely related to vandalism and bias detection, it is a more difficult problem. In this paper, we frame CoI detection as a binary classification problem and explore various features which can be used to train supervised classifiers for CoI detection on Wikipedia articles. Our experimental results show that the best F-measure achieved is 0.67 by training SVM from a combination of features including stylometric, bias and emotion features. As we are not certain that our non-CoI set does not contain any CoI articles, we have also explored the use of one-class classification for CoI detection. The results show that using stylometric features outperforms other types of features or a combination of them and gives an F-measure of 0.63. Also, while binary classifiers give higher recall values (0.81∼0.94), one-class classifier attains higher precision values (0.69∼0.74).

Keywords: Wikipedia, Conflict-of-Interest Detection, Bias, Stylometric features, One-class classification.

1. Introduction

A key feature of Wiki sites is to allow people from all over the world to add or modify articles anonymously and without consequence. This enables people with malicious intentions to use articles to promote or to discredit target products, services, organisations, or individuals.

Conflict of Interest (CoI) is defined as a situation in which a person or organisation is involved in multiple interests, financial interest, or otherwise; one of which could possibly corrupt the motivation of the individual or organisation[1]. According to Wikipedia, content on Wikipedia and other Wiki-media projects "must be written from a neutral point of view (NPOV)"[2]. NPOV refers to representing neutral and without bias all of the significant views that have been published by reliable sources. CoI editing happens when an editor contributes to Wikipedia about themselves or their relationships such as family, friends, clients, employers, and financial links, etc. Often times, CoI editing does not comply with NPOV.

CoI editing is strongly discouraged by Wikipedia as it undermines the public's confidence in it, and causes public embarrassment to the individuals being promoted. It is easy to assume that CoI is just bias; however while it is not possible for CoI to exist without bias, bias can often exist in the absence of a CoI. One's beliefs and desires can lead to biased editing, but that does not constitute a CoI.

The growth of Wikipedia makes it increasingly difficult for both Wikipedia users and administrators to manually monitor articles. Taking two example documents from our dataset, one is an article classified as CoI while the other is not:

- **CoI example:** *Kaizaad Kotwa, born in Mumbai, India, is an award winning professor and writer, actor, director, producer and designer. Currently he is a professor of theatre and film at Ohio State University in Columbus, Ohio. He recently won the Griffin Society Award for Best Professor and in 2007 was named one of the top professors in Ohio. He is the co-owner and co-Artistic Director of Poor-Box Productions, along with his mother Mahabanoo Mody-Kotwal, a famous actor, director and producer in India.*

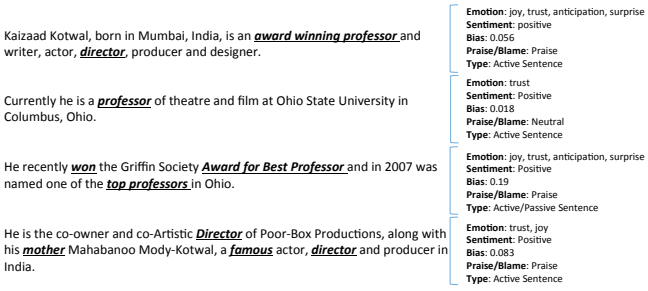
- **Non-CoI example:** *"Enrica Zunic" is the pseudonym of "Enrica Lozito", an Italian science-fiction writer. She lives and works in Turin. Her work is partly inspired by her activities with Amnesty International. In 2003 she won the Premio Italia award for science fiction.*

Using our proposed approach, a number of interesting features are identified as shown in Figure 1. It can be observed that the CoI example when compared to the non-CoI one contains more subjective sentences, bias sentences, emotion and more praise/blame expressions.

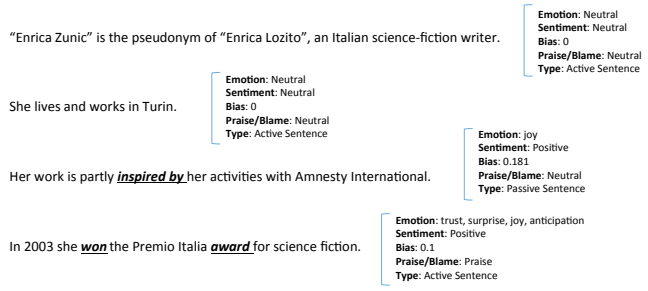
Our main aim in this work is to detect CoI articles based solely on the content of the articles without relying on any related metadata. We explore a rich set of features including stylometric features, the presentational features by focusing on the existence of Rhetorical Structure Theory's (RST's) presentational relations, various forms of language biases and implicit/explicit emotions. We then investigate using different combinations of features to train supervised binary classifiers for CoI detection. Our results show that the best result of 0.67 in F-measure is obtained when training Support Vector Machines (SVMs) from a combination of all features. Also, further combining various features with document-level representations either in the form of bag-of-words or dense representations by combining pre-trained word vectors does not bring any performance gains. As we only have the labeled CoI class, but not the non-CoI class, we have also explored the use of one-class classification for CoI detection. The results show that using stylometric features outperforms other types of features or a combination

[1] https://en.wikipedia.org/wiki/Conflict_of_interest

[2] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view



(a) CoI example article



(b) Non-CoI example article

Figure 1: Two sample documents with features identified by our approach. Words/phrases underlined in text are those which be found in an emotion or sentiment lexicon. Due to space constraint, we only show some key features here such as the emotion/sentiment label, bias score, praise/blame indicator, and sentence type.

of them. Also, one-class classifier gives higher precision values compared to binary classifiers. To the best of our knowledge, we are the first to carry out automatic CoI detection on Wikipedia articles based solely on text content. Our main contributions are summarised below:

- We have built a CoI dataset which contains 3,280 CoI articles and 3,450 non-CoI articles, which could be used in future research on CoI detection;
- We have proposed a set of features based on our research of existing work close to CoI detection and analysis of the data collected and have identified the most effective features through extensive experiments on our CoI dataset;
- We have also investigated the effectiveness of using one-class classification for CoI detection.

The problem of content-based CoI has never been investigated before. We believe that our work will inspire further development of automated systems for CoI detection based on text content.

2. Related Work

There is no prior work on CoI detection from text. But CoI is closely related to vandalism and bias. As such, we review existing work on vandalism and bias detection from textual data, with focus on Wikipedia articles.

Vandalism can be defined as any modification of content made in a cautious effort to compromise the integrity of Wikipedia (West et al., 2010). Early tools consist of bots that would label vandalism using handcrafted rules encoding heuristic vandalism patterns. Such bots include ClueBot[3], MartinBot[4], etc. These bots' typical rules were limited and some of the features they examined include: the amount of text inserted or deleted, the ratio of capital letters, and the presence of vulgarisms detected.

Chin et al. (2010) looked at constructing statistical language models of an article from its revision history. According to their approach, if inappropriate content is added to the article, then the compression level is lower than it would be for text which is similar to existing content. This approach has a drawback that it tends to label as vandalism any large addition of material, regardless of its quality, while overlooking the small additions of vandalism. The idea of using reputation systems to aid in vandalism detection was advanced in (Zeng et al., 2006; Adler and De Alfaro, 2007). West et al. (2010) applied the concept of reputations to editors and articles. They proved that the broader use of meta-data can be very effective.

Potthast et al. (2010) presented a comprehensive overview of what types of features have been employed for vandalism detection. Early approach (Potthast et al., 2008) used manual inspection to construct a feature set based on meta data and content-level properties and built a classifier using logistic regression. They achieved 83% in precision and 77% in recall. Other similar machine learning approaches for vandalism detection include those proposed in (Smets et al., 2008; Itakura and Clarke, 2009; Mola-Velasco, 2012). Harpalani et al. (2011) hypothesised that vandalism edits have unique linguistic properties in common. They based their approach on stylometric analysis of vandalism edits using probabilistic context-free grammar models. Their approach outperformed features based on shallow patterns and achieved 77% in recall.

Recasens et al. (2013) analysed real instances of human edits designed to remove bias from Wikipedia articles. The analysis uncovers two classes of bias: framing bias, such as praising or perspective-specific words link to subjectivity; and epistemological bias, related to whether propositions that are presupposed or entailed in the text are undisputedly accepted as true. They found that features based on subjectivity and sentiment lexicons are very helpful in detecting bias. Callahan and Herring (2011) examined cultural bias based on Wikipedia's NPOV policy.

Bhosale et al. (2013) presented work on detecting promotional content in Wikipedia. They looked at the content features, structural features, network features, edit history features, overall sentiment score, trigram language models and PCFG language models. They found that the stylometric features influenced results the most.

When an edit is made on Wikipedia, the editor can either register for an account or edit anonymously. When done anonymously, Wikipedia uses the IP address to identify and distinguish the article instead of a username. WikiScanner or WikiWatchdog listed "anonymous" edits related to real-world organisations. They work by comparing a list of all IP addresses that have made edits to Wikipedia with IP

[3] https://en.wikipedia.org/wiki/User:ClueBot
[4] https://en.wikipedia.org/wiki/User:MartinBot

addresses which belong to real world organisations and returning a list of "anonymously" edited articles made from the organisations' IP addresses. Although WikiScanner or WikiWatchdog can be potentially used for CoI detection, they suffer from a number of limitations, for example, they don't analyse the content itself and don't consider edits done by registered users. Also, to avoid the detection by WikiScanner or WikiWatchdog, one would simply make an edit from a IP address not belonging to a real world organisation.

3. Our Approach

We address the CoI detection problem as binary classification which determines if a given document belongs to the category of CoI or non-CoI. We make the following hypotheses:

1. Since CoI is a sub-category of the "NPOV disputes" Wikipedia category, CoI articles inherit various linguistic and stylometric characteristics from their parent Wikipedia categories including those typically found in vandalism and bias;

2. CoI articles contain more subjective sentences than non-CoI articles;

3. The presentation of content in CoI articles will tend to increase the reader's interest/regard for the subject matter;

4. Since the choice of words projects opinions and preferences, CoI articles likely contain more expressions of implicit or explicit emotions.

In this section, we explore a rich set of features to test our hypotheses above and to train supervised classifiers for CoI detection.

3.1. Stylometric Features

Stylometric features attempt to recognise patterns of style in text. These techniques have been traditionally applied to attribute authorship (Reddy et al., 2016; Stamatatos, 2009; Argamon et al., 2009), opinion mining (Panicheva et al., 2010), and forensic linguistics (Turell, 2010; Olsson and Luchjenbroers, 2013). We create a list of features selected from previous research work in vandalism and bias as mentioned in the Related Work section. Since not all features are relevant to our CoI detection task, We perform feature selection using the implementation of InfoGain and Chi-Square available in Weka[5] to eliminate insignificant features. We also include the nine universal dependency groups[6], detection of which is done using the Stanford Dependency Parser[7]. The final set of features is listed in Table 1. This set of features is relating to Hypothesis 1.

3.2. Bias Features

In (Recasens et al., 2013), two major classes of bias in Wikipedia edits have been discussed, *framing bias* and *epistemological bias*. The former is realised by subjective words or phrases linked with a particular point of view, while the latter is related to linguistic features that subtly focus on the believability of a proposition. We use the same classes of bias as discussed in (Recasens et al., 2013) and identify existence of the classes in a Wikipedia article based on a bias lexicon[8]. We also consider other words/phrases which may introduce bias as illustrated in the Wikipedia's manual of style/Words to Watch[9]. The bias features are shown in Table 2. This set of features is relating to Hypothesis 1 and 2.

3.3. Presentational Features

Rhetorical Structure Theory (RST) is a discourse theory, which offers an explanation of the coherence of texts. It provides a way to describe the relations among text and has been used to successfully analyse a variety of text types (Taboada, 2006; Taboada and Mann, 2006). In RST, presentational relations are those whose intended effect is to increase some inclination in the reader or acceptance of the content (Mann and Thompson, 1987).
We focus our work on identifying the existence of presentational relations[10] using cue words as relation signals. We use 10 presentational relations as shown in Table 3, as they increase readers' acceptance of text in one form or the other. We built a simple cue phrase detector with phrases provided in various RST research (Taboada, 2006) and relation nucleus/satellite positioning described in (Mann and Thompson, 1987). This set of features is relating to Hypothesis 3.

3.4. Emotion Features

We focus on Ekman's six basic emotions (*joy, sadness, anger, surprise, fear, disgust*) and implement both explicit and implicit emotions detection. Emotions can be expressed explicitly by using "emotion-bearing words" or implicitly without such words. For explicit emotions, we use a simple lexicon-based approach with negation handling based on a modified version of the NRC lexicon (Mohammad and Turney, 2013); and for implicit emotions, we use the rule-based approach (Udochukwu and He, 2015). In addition, we also perform polarity detection (*positive* and *negative*) using majority voting based on the lexicon matching results obtained with three sentiment lexicons, SentiWordNet (Esuli and Sebastiani, 2005), AFINN (Hansen et al., 2011) and the Subjectivity Lexicon (Wilson et al., 2005). We implement a contextual valence shifter as described in (Polanyi and Zaenen, 2006) to detect polarity change in context. Apart from emotion and polarity features, we also consider the expressions of *blame* and *praise* as additional features using the method proposed in (Orizu and He, 2016)

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://universaldependencies.org/docsv1/u/dep/index.html

[7] http://nlp.stanford.edu/software/stanford-dependencies.shtml

[8] http://www.mpi-sws.org/~cristian/Biased_language.html

[9] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch

[10] http://www.sfu.ca/rst/01intro/definitions.html

Feature Name	Description
Sentence Level	
Average_Sentence_Length	Average length of the sentences in the document
Average_Unique_Word_Count	Average # of unique words per sentence
Average_Punctuation	Average number of punctuations per sentence
Adjective_Rate	Rate of adjectives per sentence
CC_Rate	Rate of coordinating conjunctions per sentence
Pronouns_Rate	Rate of pronouns per sentence
Word_Count_Score	Total # of words / Total # of sentences
Unique_POS_per_Sentence	Rate of unique Part-of-Speech (POS) tags per sentence
Document Level	
Sentence_Count	Total # of sentences in the document
Unique_Word_Count	Total # of unique words in the document
No_of_Verbs	Total # of verbs in the document
No_of_CC	Total # of coordinating conjunctions in the document
No_of_CompAdverbs	Total # of comparative adverbs in the document
No_of_Adjectives	Total # of adjectives in the document
Special_clausal_dependents	Total # of special clausal dependents in the document
Active_Sentences	Total # of non-passive sentences
Non_core_dependents_of_clausal_predicates	Total # of non-core dependents of clausal predicates
Core_dependents_of_clausal_predicates	Total # of core dependents of clausal predicates
Noun_dependents	Total # of Noun dependents
Compounding_and_unanalyzed	Total # of Compounding and unanalyzed dependencies
Case-marking, prepositions, possessive	Total # of Case-marking, prepositions, possessive
Coordination	Total # of Coordination dependencies
Loose_joining_relations	Total # of loose joining relations
Sentence_head_and_Unspecified_dependency	Total # of Sentence head and Unspecified dependency
Complexity_Score	Text complexity score

Table 1: Stylometric features.

Bias	Subtypes
Epistemological	Factive verbs / Entailments / Assertives / Hedges
Framing	Subjective terms / Intensifiers
Others	Puffery / Contentious labels / Unsupported attributions / Expressions of doubt / Editorialising

Table 2: Bias features and subtypes.

Relation Name	Intention of W
Antithesis	R's positive regard for N is increased
Background	R's ability to comprehend N increases
Concession	R's positive regard for N is increased
Enablement	R's potential ability to perform the action in N increases
Evidence	R's belief of N is increased
Justify	R's readiness to accept W's right to present N is increased
Motivation	R's desire to perform action in N is increased
Preparation	R is more ready, interested or oriented for reading N

Table 3: The 10 presentational relations used (N stands for nucleus, R for reader and W for writer).

for detection. This set of features is relating to Hypothesis 4.

4. Experiments

4.1. Data

We construct our dataset by collecting 4,050 articles from Wikipedia which have been categorised as conflict of interest (CoI) items[11]. This CoI category is a sub-category of "NPOV disputes". Wikipedia encourages its editors to pick an article from this category and decide whether it meets its notability policy[12]. If one believes the article should be kept, he/she needs to review the text to ensure that it complies with NPOV. This human categorisation of Wikipedia articles will be our basis for evaluating our results.

In order to build a dataset containing both CoI and non-CoI articles, for each CoI article, we randomly select non-CoI articles from its first associated Wikipedia category. For example, a CoI article might be associated with two categories, "1932 births" and "Living people". We randomly select a non-CoI article from the category "1932 births". This resulted in a total of 4,600 non-CoI articles selected from over 100 Wikipedia categories. We have considered various criteria for the selection of non-CoI articles such as age of article, number of views, editor information. We found that identifying a threshold on these meta-data that cuts across the various categories and sectors would require a fine-tooth comb. For example, an article maybe older but has fewer views than a newer article OR articles from a par-

[11] https://en.wikipedia.org/wiki/Category:Wikipedia_articles_with_possible_conflicts_of_interest

[12] https://en.wikipedia.org/wiki/Wikipedia:Notability

ticular category may have more views than other categories. As a result, we chose the random selection approach as long as the article was from the same category as a CoI article and did not belong to CoI disputes category. We focus on the article content as our means of classification and ignore the meta information provided by Wikipedia such as the editor(s) of a Wikipedia edit, time and date of creation, associated IP address, etc.

4.2. Preprocessing

We pre-process the dataset by removing the top 1% articles and the lower 5% of the articles based on the document length. This reduces the total number of document to 3,280 CoI articles and 3,450 non-CoI articles. The vocabulary size for the dataset is 52,302. We then carry out sentence splitting and tokenisation, stopword removal, stemming and remove words occurred less than ten times. For implicit emotion detection and blame/praise detection, we also perform part-of-speech (POS) tagging using the Stanford POS Tagger[13], word sense disambiguation (WSD) using the classic Lesk algorithm for WSD in NLTK[14], and dependency parsing using the Stanford Dependency Parser. To represent documents, apart from the commonly used bag-of-words approach, we also consider using doc2vec (Le and Mikolov, 2014) which modifies the word2vec algorithm (Mikolov et al., 2013) for unsupervised learning of continuous representations for larger blocks of text, such as sentences, paragraphs or entire documents. Recent work in the area of NLP has shown it to be a strong alternative for both bag-of-words and bag-of-n-grams models. We use Gensim[15] which has an implementation of doc2vec. We ignore words occurred less than 10 times and generate a vector representation of each article using the pre-trained vectors from the Google News dataset[16] with about 100 billion words, 300-dimensional vectors. The size of the context window we use is 3 before and after the predicted word. The final generated document vectors have 100 dimensions.

4.3. Feature Selection

Here we aim to identify the features that are mostly useful for prediction of CoI. We use Correlation-based Feature Subset Selection (CFS) and Information Gain Ratio (IGR) to rank features on all our feature sets from the training set and merge the top 15 features as listed in Table 4. Most of the top features are Stylometric features (74%) followed by the Emotion (21%) and Bias (5%) features. We also found that no Presentational features appear in the top 15 positions. The feature selection results indicate that stylometric features are very important in determining whether an article should be classified as CoI. Among various emotion features, *Blame*, *Praise*, *Polarity_Score* and *Suprise* seem more important than others. The *Bias_Score* is also relevant, but less important compared to many Stylometric or some Emotion features. Presentational features do not seem to contribute much to CoI detection.

Feature	Set	Description
Blame	*Emotion*	Total # of expressions of "Blame"
Praise	*Emotion*	Total # of expressions of "Praise"
Polarity_Score	*Emotion*	Aggregated polarity score of the document
Surprise	*Emotion*	Total # of expressions of "Surprise"
Active_Sentences	*Stylometric*	Total # of non-passive sentences
Non_core_dependents _of_clausal_predicates	*Stylometric*	Total # of non-core dependents of clausal predicates
Average_Sentence_Length	*Stylometric*	Average length of sentences in the document
Average_Unique_Word_Count	*Stylometric*	Average # of unique words per sentence
No_of_CC	*Stylometric*	Total # of coordinating conjunctions in the document
CC_Rate	*Stylometric*	Rate of coordinating conjunctions per sentence
Adjective_Rate	*Stylometric*	Rate of adjectives per sentence
Pronouns_Rate	*Stylometric*	Rate of pronouns per sentence
Sentence_Count	*Stylometric*	Total # of sentences in the document
Coordination	*Stylometric*	Total # of Coordination dependencies
Word_Count_Score	*Stylometric*	Total # of words / Total # of sentences
Unique_POS_per _Sentence	*Stylometric*	Rate of unique Part-of-Speech (POS) tags per sentence
Complexity_Score	*Stylometric*	Text complexity score
Special_clausal_dependents	*Stylometric*	Total # of special clausal dependents
Bias_Score	*Bias*	Aggregated bias score of the document

Table 4: Merged features from the feature selection results from Correlation-based Feature Subset Selection (CFS) and Information Gain Ratio (IGR).

4.4. Binary Classification Results

We train supervised classifiers including Support Vector Machines (SVMs), Maximum Entropy (MaxEnt) and Naïve Bayes (NB) using various feature sets and different combinations of them. Ten-fold cross validation is used and the results are averaged over 10 such runs.

We can observe from Table 5 that among the four feature sets, *Stylometric* gives the best performance followed by *Emotion*. This is consistent with our feature selection results discussed in Section 4.3.. It also confirms our hypothesis that the writing styles of editors of CoI articles are similar. *Bias* and *Presentational* features appear to be less useful. This shows that CoI is more than just bias. Presentational features had no member appeared in the top 20 features ranked by CFS or IGR. Although SVM or MaxEnt trained from *Presentational* or *Bias* features give much worse results compared to other feature sets, NB trained from these two types of features sets performs only slightly worse than trained from *Stylometric* or *Emotion* features.

We have also tried with combinations of different features sets. For both SVM and MaxEnt, the best performance is given by *All features*. SVM achieves much higher recall than precision with an overall F-measure of 0.67. MaxEnt gives more balanced precision and recall values, but with slightly worse F-measure compared to SVM. We also notice that using *Best features* as listed in Table 4 does not lead to improved performance for SVM or MaxEnt. However, the *Best features* set boosts the recall value to 0.94 for NB, although it only gives the precision value of 0.51.

We have next experimented with document representations using Bag-of-Words (BOW) weighted by TFIDF or doc2vec, and a combination of BOW or doc2vec with various feature sets. But they do not give any improvements, showing that CoI classification is not relevant to words presented in documents. Due to the page limit, we do not report the results here.

[13] http://nlp.stanford.edu/software/tagger.shtml

[14] http://www.nltk.org/howto/wsd.html

[15] http://radimrehurek.com/gensim/

[16] https://code.google.com/archive/p/word2vec/

Feature Sets	SVM			MaxEnt			NB		
	P	R	F	P	R	F	P	R	F
Stylometric	0.57	0.74	0.64	0.60	0.61	0.60	0.50	0.91	0.65
Presentational	0.07	0.06	0.06	0.57	0.36	0.44	0.49	0.90	0.63
Bias	0.18	0.16	0.17	0.57	0.26	0.35	0.48	0.91	0.63
Emotion	0.26	0.16	0.24	0.56	0.39	0.46	0.49	0.92	0.64
Stylometric+Emotion	0.57	0.73	0.64	0.60	0.61	0.61	0.50	0.91	0.65
Stylometric+Emotion+Bias	0.57	0.73	0.64	0.61	0.61	0.61	0.50	0.91	0.65
All features	0.58	**0.81**	**0.67**	**0.64**	**0.67**	**0.66**	**0.62**	0.51	0.56
Best features	**0.61**	0.30	0.40	0.61	0.65	0.63	0.51	**0.94**	**0.66**

Table 5: Conflict-of-Interest (CoI) detection results in Precision, Recall and F-measure using SVM, MaxEnt and NB with various feature sets.

4.5. One-Class Classification

In Section 4.4., we train supervised classifiers from a dataset containing both CoI and non-CoI documents for binary classification. One problem we encountered is that there is no-degree of assurance that the items in our non-CoI category are purely non-CoI documents, as they where merely selected randomly from the same Wikipedia categories as CoI articles, with no concrete certainty that they are all non-CoI. Our problem could be potentially solved by one-class classification (Manevitz and Yousef, 2001; Khan and Madden, 2009), in which one of the target class is well represented by instances in the training data with little or no other class present. The problem of One-class classification is harder than the problem of conventional classification as a result of the one-sided nature of the dataset. One-class classification makes it difficult to decide which attributes should be used to best separate target and non-target (i.e., CoI and non-CoI in our case).

In (Schölkopf et al., 2001), adapting SVM to the one-class classification problem has been proposed. Essentially, the input data are first mapped into a high dimensional feature space via a kernel. The origin is considered as the only member of the second class. Then the algorithm iteratively finds the maximal margin hyperplane which best separates the training data from the origin. In our experiments here, we used one-class SVM implementation in the LIBSVM[17] with default parameters.

Feature Sets	Precision	Recall	F-Measure
Stylometric	**0.74**	**0.55**	**0.63**
Presentational	0.69	0.52	0.59
Bias	0.72	0.54	0.62
Emotion	0.73	0.55	0.62
All Features	0.72	0.53	0.61
Best features	0.73	0.54	0.62

Table 6: CoI detection results using one-class classification.

Table 6 shows the CoI detection results using one-class classification by 10-fold cross validation trained on the CoI-related documents only. It can be observed that using *Stylometric* features gives the best results compared to other feature sets although the improvement in F-measure compared to the *Bias* or *Emotion* features is only marginal. We also notice that the precision values, which are in the rage of 0.69 to 0.74, are much higher than those achieved based on binary classification where the typical precision values are between 0.58 and 0.64. However, the recall values are lower (0.52∼0.55 cf. 0.81∼0.94). This shows that if we aim to achieve high recall values for CoI detection, then binary classification should be used. However, if high precision values are more desirable, then one-class classification should be used instead.

[17] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

4.6. Comparison with an Existing Approach to Vandalism Detection

There is no prior approach to content-based CoI detection from Wikipedia. Existing work to bias or vandalism detection often made use of metadata such as anonymity, edit frequency, author reputation, etc., and performed classification at the sentence-level. As we do not have the relevant metadata available and there are no sentence-level annotations in our dataset, directly comparing our approach with existing work is difficult. Nevertheless, we re-implemented an approach proposed in (Mola-Velasco, 2012) in which their best F-measure and AUC were achieved using LogitBoost and Random Forest, respectively, ranking in the first place of the PAN'10 Wikipedia vandalism detection task (Potthast et al., 2010). Since we do not have edit histories available, we exclude features relating to edit histories and only extract other stylometric features and features analogous to vulgarism frequency and vulgarism impact and train LogitBoost for 500 iterations. The results in comparison to our best ones are listed in Table 7. It can be observed that both our binary and one-class classifiers outperform LogitBoost with the performance gain in F-measure ranging from 6% to 10%.

Method	Precision	Recall	F-Measure
LogitBoost	0.56	0.58	0.57
SVM (binary)	**0.74**	0.55	0.63
SVM (one-class)	0.58	**0.81**	**0.67**

Table 7: Comparison with an existing approach to vandalism detection.

4.7. Discussion

Our finding of the importance of stylometric features confirms our original hypothesis in Section 3. that CoI will inherit linguistic and stylometric features from its parent Wikipedia Category. But our hypothesis that presentation relations would affect CoI was not supported by our experimental results. We found that our hypothesis on CoI articles being more subjective holds true based on the experiment results. Also, the hypothesis that CoI articles contain more expressions of implicit and explicit emotions is also supported from our experimental results.

Our feature selection results show that *Blame*, *Praise* and *Polarity_Score* are discriminative features for the CoI class as they are ranked in the top 3 positions by CFS. However, in binary classification results, using features from the *Emotion* category gives worse results compare to the *Stylometric* category, although it outperforms both *Presentational* and *Bias* categories. The same observation holds for one-class classification. Using *Stylometric* features consistently outperform other feature sets for both binary and one-class classification. Also, it seems that articles with a higher rate of coordinating conjunctions and adjectives per sentence have a higher chance of belonging to the CoI category.

5. Conclusions and Future Work

The work presented here tackles a unique problem for the automatic detection of Conflict of Interest (CoI) articles in Wikipedia entries based on the content of the articles. We have shown that the CoI detection task is a complex problem but with carefully engineered feature sets, it is possible to identify CoI articles with an F-measure of 0.67 using SVM. We have also found that out of four different sets of features, *Stylometric* features help the most with CoI detection. In addition to binary classification, we have experimented with one-class classification and shown that while binary classification gives higher recall values, one-class classification attains higher precision values.

In future work, we intend to explore other types of features extracted from metadata of Wikipedia articles such as editors' information, editing history and associated IP addresses, and evaluate their impact on the performance of CoI detection. It is possible that articles in different Wikipedia categories might follow different writing styles (e.g., Wikipedia entries about people and about organisations). One possible direction is to build category-specific classifiers for CoI detection. Finally, to avoid expensive feature engineering, it is possible to learn feature representations and classifiers simultaneously by investigating various deep learning architectures.

6. Bibliographical References

Adler, B. T. and De Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, pages 261–270.

Adler, B. T., De Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 277–288.

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Bhosale, S., Vinicombe, H., and Mooney, R. J. (2013). Detecting promotional content in wikipedia. In *EMNLP*, pages 1851–1857.

Callahan, E. S. and Herring, S. C. (2011). Cultural bias in wikipedia content on famous persons. *Journal of the Association for Information Science and Technology*, 62(10):1899–1915.

Chin, S.-C., Street, W. N., Srinivasan, P., and Eichmann, D. (2010). Detecting wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th workshop on Information credibility*, pages 3–10.

Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624.

Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., and Etter, M. (2011). Good friends, bad news-affect and virality in twitter. *Future information technology*, pages 34–43.

Harpalani, M., Hart, M., Singh, S., Johnson, R., and Choi, Y. (2011). Language of vandalism: Improving wikipedia vandalism detection via stylometric analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 83–88.

Itakura, K. Y. and Clarke, C. L. (2009). Using dynamic markov compression to detect vandalism in the wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 822–823.

Khan, S. S. and Madden, M. G. (2009). A survey of recent trends in one class classification. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 188–197.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Manevitz, L. M. and Yousef, M. (2001). One-class svms for document classification. *Journal of Machine Learning Research*, 2(Dec):139–154.

Mann, W. C. and Thompson, S. A. (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).

Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Mola-Velasco, S. M. (2012). Wikipedia vandalism detection through machine learning: Feature review and new proposals: Lab report for pan at clef 2010. *arXiv preprint arXiv:1210.5560*.

Olsson, J. and Luchjenbroers, J. (2013). *Forensic linguistics*. A&C Black.

Orizu, U. and He, Y. (2016). Detecting expressions of blame or praise in text. In *the 10th edition of the Language Resources and Evaluation Conference*, Portoro, Slovenia, May, 2016.

Panicheva, P., Cardiff, J., and Rosso, P. (2010). Personal sense and idiolect: combining authorship attribution and opinion analysis. In *the 7th International Conference on Language Resources and Evaluation*, Malta, 2010.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. *Computing Attitude and Affect in Text*, 20:1–10.

Potthast, M., Stein, B., and Gerling, R. (2008). Automatic vandalism detection in wikipedia. In *European Conference on Information Retrieval*, pages 663–668.

Potthast, M., Stein, B., and Holfeld, T. (2010). Overview of the 1st international competition on wikipedia vandalism detection. In *CLEF (Notebook Papers/LABs/Workshops)*.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659.

Reddy, T. R., Vardhan, B. V., and Reddy, P. V. (2016). A survey on authorship profiling techniques. *International Journal of Applied Engineering Research*, 11(5):3092–3102.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Smets, K., Goethals, B., and Verdonk, B. (2008). Automatic vandalism detection in wikipedia: Towards a machine learning approach. In *AAAI workshop on Wikipedia and artificial intelligence: An Evolving Synergy*, pages 43–48.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556.

Taboada, M. and Mann, W. C. (2006). Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations. *Journal of pragmatics*, 38(4):567–592.

Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech, Language & the Law*, 17(2).

Udochukwu, O. and He, Y. (2015). A rule-based approach to implicit emotion detection in text. In *International Conference on Applications of Natural Language to Information Systems*, pages 197–203.

West, A. G., Kannan, S., and Lee, I. (2010). Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata? In *Proceedings of the Third European Workshop on System Security*, pages 22–28.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Zeng, H., Alhossaini, M. A., Ding, L., Fikes, R., and McGuinness, D. L. (2006). Computing trust from revision history. Technical report, Stanford Univ Ca Knowledge Systems LAB.