EXPLORATORY ANALYSIS OF PRESS ARTICLES ON KENYAN ELECTIONS: A DATA MINING APPROACH

Senja Pollak
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: senja.pollak@ijs.si

ABSTRACT

This paper investigates the utility of applying data mining techniques to media analysis, more specifically, to the analysis of a corpus of articles covering the 2007 Kenyan election and post-election crisis, aimed at capturing the differences between local (Kenyan) and Western (British) newspaper articles. Having formulated this task as a binary classification problem, we have succeeded to reveal interesting phenomena in the data using data/text classification methods and class association rules, opting for techniques where interpretability of results prevails over their accuracy.

1 INTRODUCTION

Media analysis has been a topic of several studies, including a recent media analysis study performed by Fortuna et al. (2008). Text classification methods prove to be a useful vehicle e.g., for different newspaper article classification tasks, such as article genre, topic or author classification.

One of the starting points of this paper is the theory of pragmatics, which pays a lot of attention to choice-making in language use (Verschueren, 1999). Lexical, syntactic or discursive choices are significant: the use of words, syntactic structures, modality markers, etc. as well as absence of their use is always meaningful.

Our approach to text classification is formulated as a binary classification task, aimed at distinguishing between the articles from Kenyan newspaper *Daily Nation* and British newspaper *The Independent* (a forthcoming, more extensive experimental study will take a larger set of articles from different Kenyan and European/American newspapers, and will – like in this study – aim at distinguishing between two selected classes: *local* and *Western*). The starting hypothesis of this work is that news coverage of Kenyan events is not the same in the local and in the international (Western) media. Using data mining/machine learning techniques, the main goal of our work is to explore how some of the lexical choices differ in the Western and local media.

The structure of this paper is as follows. In Section 2 we present the data. Section 3 outlines the text classification and machine learning techniques used in our analysis. Section 4 presents selected results of the analysis. Section 5 presents the conclusions and plans for further work on a larger selection of 464 newspaper articles.

2 DATA

This section presents the data (presented in Section 2.1), data cleaning (2.2) and data representation (2.3).

2.1 Data description

Originally, the corpus was collected as part of the project *Intertextuality and Flows of Information* in the field of pragmatics. The collected corpus consists of articles from six different newspapers (Kenyan, British and American) in English language, covering the Kenyan election and post-election crisis between December 2007 and April 2008.

The Kenyan presidential and parliamentary elections were held on the 27th of December 2007. The two main candidates were the incumbent President Mwai Kibaki and the opposition presidential candidate Raila Odinga. Kibaki is a member of the traditionally dominant Kikuyu ethnic group and Odinga is a member of the Luo ethnic group. Kibaki was declared the winner and sworn in despite the opposition leader's claims of victory. The election was followed by violence and conflicts.

For the first experiments, which are the focus of this paper, we had only a limited set of articles at our disposal. Our data set consists of 72 articles from only two newspapers: 36 articles form *The Independent* for Western media (WE) and 36 from *Daily Nation* for local media (LO). This subset was used in the reported study and will be followed by experiments on a selection of a larger number of articles.

2.2 Cleaning the data

Since our aim is to better understand the way of reporting on the same event by two different newspapers, we had to remove all information that could be distinctive for the two classes, but not important for our work. To illustrate, newspapers have normally only few journalists covering Kenya events, so if not removed, the author's name could easily be selected as a distinguishing feature. Therefore, we removed meta-information such as newspaper source, authors of articles, dates of publication, photographers, mails of authors, types of articles, etc. For this purpose, we made scripts in Perl and used only the remaining relevant data for document classification (titles, text and photo descriptions).

2.3 Data representation

Each of the two classes, LO (local) and WE (Western), contains 36 instances (articles). The class is a nominal attribute.

For data representation we had to select the attributes, and represent the articles with feature vectors. We selected word unigrams (W1) and word bigrams (W2) as attributes. The attributes have a numeric value, calculated on the basis of term frequency (*tf*). Since having all the attributes would result in too large feature vectors, attribute selection of best 500 attributes ranked by *chi square* values was first performed. This feature selection and transformation to feature vectors was done in TACTiCS[1]. We also used a binary representation of this selection of attributes (1 - word is present in the document, and 0 – word is not present in the document) as well as concatenation of word unigrams and word bigrams. To summarize, we experimented with the following feature sets: W1 (word unigrams weighted by term frequency), W2 (word bigrams weighted by term frequency), W1-bin (word unigrams transformed to binary representation), W2-bin (word bigrams transformed to binary representation), W1W2 (concatenated W1 and W2 feature sets: 500 selected word unigrams followed by 500 selected word bigrams).

3 TEXT CLASSIFICATION AND MACHINE LEARNING ALGORITHMS USED

This section starts with a theoretical definition of text classification and situates our task within this framework in Section 3.1, while Section 3.2 presents the tools and motivates our choice of machine learning techniques used.

3.1 Text classification

Text classification can be defined as "automatic assignment of documents to a predefined set of categories" (Sebastiani, 2002), and learning of text classifiers (classification models) can be performed by supervised machine learning.

Text categorization is the task of assigning a binary value (T or F) to each pair $<d_j, c_i> \in D \times C$, where D is a domain of documents and $C = \{c_1,...,c_{|C|}\}$ is a set of predefined categories. In our case C = {WE, LO}, WE meaning Western media and LO meaning local Kenyan media.

Since we want exactly one category to be assigned to each $d_j \in D$, and we have only two complementary categories $|C|=2$, our case is an example of binary text classification, a special case of a general N class single-label (or non-overlapping categories) case.

3.2 Algorithms used

The choice of symbolic data mining algorithms from the Weka 3.6.0 environment[2] was motivated by the need for ensuring the interpretability of results, possibly at a cost of not achieving the highest possible accuracy. For this reason, we used decision tree and decision rule classifiers and not better performing algorithms such as support vector machines or nearest neighbours algorithms. We used the following algorithms available in Weka: J48 for learning decision trees (Quinlan, 1986 and 1993), 1R for decision rules (Holte, 1993), JRip for decision rules (Cohen, 1995), PRISM for decision rules (Cendrowska, 1987), and PART for decision list learning (Frank and Witten, 1998). An experiment was done also by learning of class association rules with Predictive Apriori (Scheffer, 2001).

4 RESULS, EVALUTATION AND INTERPRETATION

This section is divided into two parts. Section 4.1 presents some results of algorithms applied to the entire data set (72 instances) without performing accuracy evaluation. In this setting we used no pruning, trying to build models and patterns that describe the given data set. This part is called exploratory data analysis (descriptive analysis). In Section 4.2 we present the results of predictive/classificatory data analysis on which evaluation was done with 10-fold cross-validation.

4.1 Exploratory analysis

We present the construction of models induced from the whole data set. The classification accuracy of these models, if evaluated on the training data itself, can be interpreted as an upper bound for the model's performance of new, unseen data (nearly always 100%) (Witten and Frank, 2005).

Results of JRip are presented in Table 1. We see that we get some interesting descriptions of the corpus. We present the rules that cover the Western articles (where local articles are treated as 'else'). We must keep in mind for the interpretation that in learning of these rules, the examples covered by the currently constructed rule are excluded from the data in the next rule construction iteration, hence each rule below covers only the examples not yet covered by previous rules. We present the analysis of two rules, which show how the information obtained with text mining tools could be useful for further discourse analysis. We present the results on the binary W1-bin feature set.

Table 1: JRip rules induced from the W1-bin feature set for class WE (minimum number of object=1, without pruning)[3].

(raila = 1) and (tribe = 1) => class=WE (11.0/0.0)
(raila = 1) and (go = 1) and (next= 0) => class=WE (6.0/0.0)
(may = 1) and (new = 1) => class=WE (4.0/0.0)
(union = 1) and (john = 1) => class=WE (3.0/0.0)
(national = 0) and (major = 1) and (nations = 0) => class=WE (4.0/0.0)
(rather = 1) and (real = 0) => class=WE (2.0/0.0)
(by = 0) and (its = 0) => class=WE (2.0/0.0)
(could = 1) and (raila = 0) and (calm = 0) and (running = 0) => class=WE (3.0/0.0)
(alone= 1) and (emergency = 1) => class=WE (1.0/0.0)
=> class=LO (36.0/0.0)

Rule 1 that covers 11 out of 36 Western articles says that if words *Raila* and *tribe* are used, the article belongs to the Western class. *Raila Odinga* is the Kenyan presidential candidate that lost the election. The choice of word *tribe* is very interesting for the analysis. *Tribe* is a very ideologically marked word. If we talk about tribes, tribal wars and conflicts, it can be a pejorative, 'savage'

[1] *TACTiCS:* http://www.cnts.ua.ac.be/stylometry/demo.html.

[2] *Weka:* Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/

[3] The first number in the parenthesis after the class means the number of examples correctly covered by the rule; the second number means false positive examples and is in our case always 0.0.

description of African situation. Talking about the presidential candidate and at the same time of tribal division exclusively in the Western media should be further analysed. As explained earlier, the two main candidates *Raila Odinga* and *Kibaki* are from two different ethic groups, but emphasizing the tribal aspect is an ideological choice of Western media.

Rule 4 is also interesting for the analysis, where we must keep in mind that the rules cover the examples that have not yet been covered by the preceding rules. If we check the articles where words *union* and *John* appear at the same time, we see that Rule 4 covers the articles, which refer to *John Kufuor.*

- *John Kufuor, the head of the African Union...*
- *John Kufuor, the President of Ghana who is the current chairman of the African Union,...*

The next set of rules was obtained with the PRISM algorithm. PRISM generates only rules with 100% accuracy. We obtained a large set of rules perfectly covering the entire corpus. In Table 2 we present just first three rules describing the *Western* and the *local* class. Since PRISM selects only 100% accurate rules, it presents a different view on our data and the rules are much more specific.

Table 2: Selection of PRISM rules on the W1-bin feature set.

If briefly = 1 then WE
If tribe = 1 and rest = 0 then WE
If challenger = 1 and continue = 0 then WE
If running = 1 and national = 0 then LO
If raila = 0 and go = 1 then LO
If kericho = 1 and challenger = 0 then LO

We made experiments also with association rules (Predictive Apriori). For applying association rules, we first performed feature set reduction (choosing a smaller number of attributes from initially 500 attributes selected). We performed attribute filtering with different filtering methods and obtained new feature sets of binary word unigrams. We present the association rules on the feature set obtained by previous selection of 10 attributes by Relief (with kNN set to 10). The selected features are: *emergency, tour, described, tribe, raila, sell, sharing, cancel, union, could.* With Predicitve Apriori we can mine class association rules and the measure used is predictive accuracy, combining confidence and support into a single measure (Scheffer, 2001). We searched for the best five rules. For the feature set selected on W1-bin by Relief, the induced class association rules are presented in Table 3.

Table 3: First five class association rules on the W1-bin feature set.

1. tribe=1 raila=1 11 ==> class=WE 11 acc:(0.9923)
2. union=1 tribe=1 4 ==> class=WE 4 acc:(0.96755)
3. tribe=1 described=1 4 ==> class=WE 4 acc:(0.96755)
4. tribe=1 emergency=1 2 ==> class=WE 2 acc:(0.92507)
5. tribe=1 12 ==> class=WE 11 acc:(0.9175)

We see again that *tribe* is the most characteristic word appearing in class Western in combination with words like *raila, union, emergency* or appearing alone.

4.2 Predictive analysis

In this section we present the classification models on which we measure the accuracy. The evaluation criterion was the percentage of correctly classified instances. For model testing we used 10-fold cross-validation, where nine folds are used for building the classifier, one fold is used for testing, and the average accuracy is computed by repeating this action 10 times.

We made a set of experiments using different feature sets. In the majority of cases the results show quite a low accuracy and a high standard deviation. We only present the most interesting results, so not all the algorithms or feature sets are covered. Results of the experiments are presented in Table 5.

1R selects only the most important rule. Surprisingly, this is one of the best performing algorithms on our data set. We present the result on the word unigrams feature set. For W1 the accuracy is 69.64% and the chosen feature is *Raila.* For interpretation purposes the result on W1-bin is better/simpler, even if the accuracy is lower (61.43%):

```
Raila:
        0       -> LO
        1       -> WE
```

The next set of experiments was done with J48. Among the parameters that we can choose in J48 for decision tree pruning we used the minimal number of objects in the leaves. Without additional feature selection, the best results were obtained by the tree built on the concatenated W1W2 feature set and the minimal number of objects set to 4 (accuracy 58.9% but high standard variation). We can improve the obtained results by first applying a wrapper feature subset selection method. Wrapper approaches are being tuned to the learning algorithm being used.

We chose the same settings for the wrapper feature selection and for classifier learning. The minimum number of objects was set to two. The classifier has achieved 80.56% accuracy (this is the best accuracy achieved in our experiments on the small dataset of 72 documents) and the model is presented in Table 4.

Table 4: Decision tree for the W1W2 feature set, with previous wrapper feature subset selection (accuracy: 80.56%).

```
tribe <= 0
|  political/leaders <= 0
|  |  what <= 0.001479
|  |  |  pledged<= 0
|  |  |  |  forces <= 0
|  |  |  |  |  what <= 0.000516
|  |  |  |  |  |  it/had<= 0
|  |  |  |  |  |  |  union <= 0.002273: LO (30.0/7.0)
|  |  |  |  |  |  |  union > 0.002273: WE (4.0)
|  |  |  |  |  |  it/had> 0: WE (3.0)
|  |  |  |  |  what > 0.000516: WE (5.0)
|  |  |  |  forces > 0: WE (3.0)
|  |  |  pledged> 0: WE (3.0)
|  |  what > 0.001479: LO (8.0)
|  political/leaders > 0: LO (4.0)
tribe > 0: WE (12.0/1.0)
```

At the first node we find the word *tribe* like in the majority of the above presented examples. The presence of *tribe* results in leaf WE, while its absence leads us to the

Table 5: Results of 10-fold cross-validation model testing.

Algorithms	One R	J48-M2	J48-M4	JRip-N2	JRip-N4	.PART '-M 2	.PART '-M 4
Feature set	*Classification accuracy (±st.dev.)*						
W1	69.64(16.60)	47.86(25.83)	43.75(17.84)	66.96(25.49)	65.18(21.58)	36.25(18.82)	42.32(18.48)
W1_BIN	61.43(13.13)	37.14(18.71)	47.32(15.37)	59.11(24.66)	59.11(24.66)	31.96(16.42)	57.50(21.46)
W2	68.04(11.56)	27.86(13.15)	33.57(18.09)	60.00(19.76)	24.11(15.27)	60.00(19.76)	38.93(14.79)
W2_BIN	63.93(11.69)	29.64(17.48)	49.82(12.38)	62.50(13.36)	63.93(11.69)	24.82(14.93)	36.25(19.38)
W1W2	62.32(12.04)	36.96(21.72)	58.93(23.88)	56.96(19.28)	57.14(22.08)	31.25(14.99)	51.79(18.39)
W1W2_wrapperJ48	64.11(15.80)	**80.54(11.92)**	70.71(12.69)	50.36(22.87)	51.79(23.76)	71.07(16.01)	70.71(12.69)
W1W2_wrapperPART	69.82(15.28)	70.77(15.35)	70.59(14.36)	60.48(16.42)	62.93(16.62)	72.36(14.28)	70.89(13.74)

second node with an interesting word bigram (*political leaders*). The presence of this bigram is the indicator for a Kenyan newspaper.

Another model (presented in Table 6) was obtained with PART with previous wrapper feature selection.

Table 6: PART decision list for the W1W2 feature set, with previous wrapper feature subset selection (accuracy: 86.1%).

```
tribe <= 0 AND real<= 0.001582 AND running <= 0 AND emergency <=
0.000557 AND and/the > 0 AND  and/the <= 0.003215: WE (24.0/4.0)
tribe <= 0: LO (36.0/5.0)
: WE (12.0/1.0)
```

5 CONCLUSIONS AND FUTURE WORK

We can see that given a 50%-50% prior class distribution, the used machine learning algorithms do not result in very high classification accuracy on a small dataset of 72 articles. As indicated by the initial experiments on a larger data set (464 instances, accuracy around 90%), low accuracy is mainly due to a small data set available for this study. Nevertheless, machine learning algorithms do provide new insights leading to improved understanding of lexical choices.

We utilized data mining techniques in order to find out which word unigrams and word bigrams can be interpreted as the distinguishing words between the *Daily Nation* Kenyan newspaper and *The Independent.* We observed that better results are obtained on numeric feature sets (based on term frequency) than on binary ones. However, binary representation enables much easier interpretation. The best results were obtained when classification was combined with wrapper feature selection.

From the content point of view, the main selected feature was *tribe*, that appears in Western newspaper *The Independent* and not in Kenyan *Daily Nation.* This is an important finding, because *tribe* is not an ideologically neutral term, but is frequently used for stereotyping the African situation (Ray, 2008) and promoting "a myth of primitive African timelessness" (Lowe et al. 2008).

In further work we will repeat and extend the experiments by analysing 232 articles for each class (local and Western) from six different newspapers (where initial experiments indicate substantially increased accuracies of induced models). We plan to experiment also with lemmatized or stemmed feature sets, with previous stopword removing and using also syntactic features.

ACKNOWLEDGMENTS. The data was obtained from the Antwerp Center of Pragmatics, University of Antwerp. The corpus was collected by L. Michiels and R. Coesemans as a part of the project Intertextuality and Flows of Information, led by Prof. J. Verschueren. I am grateful to my supervisor, prof. W. Daelemans, as well as to Kim Luyckx for her help in transforming the data into feature vector format using system TACTiCS, developed at CNTS.

REFERENCES

Cendrowska, J. (1987). PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 27(4), p. 349-370.

Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, p.115–123.

Fortuna, B., Galleguillos, C. and Cristianini, N. (2008). Detecting the bias in media with statistical methods. In *Text Mining: Theory and Applications*, Taylor and Francis Publisher, London.

Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. In *Proceedings of the 15th International Conference on Machine Learning*, p. 144-151.

Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 6 (11), p. 63-91.

Lowe C., Brimah, T., Marsh P.-A., Minter, W. and Muyangwa, M. (1997, updated 2008). Talking about "tribe" moving from stereotypes to analysis. http://www.africaaction.org/bp/documents/TalkingaboutTribeFeb2008Update_001.pdf, last accessed 18. 8. 2009.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1, p. 81–106.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco.

Ray, C. (2008). How the word 'tribe' stereotypes Africa. New African 471. p. 8-9.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34 (1), p. 1-47.

Scheffer, T. (2001). Finding association rules that trade support optimally against confidence. *In Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, p. 424–435.

Verschueren, J. (1999). *Understanding Pragmatics*, Arnold (Understanding Language Series), London.

Witten, I. H. and Frank, E. (2005). *Data Mining Practical Machine Learning Tools and Techniques*, 2nd ed., Elsevier, San Francisco.