Classifying Italian newspaper text: news or editorial?

Pietro Totis
Università degli Studi di Udine
totis.pietro@spes.uniud.it

Manfred Stede
Applied Computational Linguistics
University of Potsdam, Germany
stede@uni-potsdam.de

Abstract

English. We present a text classifier that can distinguish Italian news stories from editorials. Inspired by earlier work on English, we built a suitable train/test corpus and implemented a range of features, which can predict the distinction with an accuracy of 89,12%. As demonstrated by the earlier work, such a feature-based approach outperforms simple bag-of-words models when being transferred to new domains. We argue that the technique can also be used to distinguish opinionated from non-opinionated text outside of the realm of newspapers.

Italiano. *Presentiamo una tecnica per la classificazione di articoli di giornale in italiano come articoli di cronaca oppure editoriali. Ispirandoci a precedenti pubblicazioni riguardanti la lingua inglese, abbiamo costruito un corpus adatto allo scopo e selezionato un insieme di caratteristiche testuali in grado di distinguere il genere con un accuratezza dell' 89,12%. Come dimostrato dai lavori precedenti, questo approccio basato sulle proprietà del testo mostra risultati migliori rispetto ad altri quando trasferito a nuovi argomenti. Riteniamo inoltre che questa tecnica possa essere usata con successo anche in contesti diversi dagli articoli di giornale per distinguere testi contenenti opinioni dell'autore e non.*

1 Introduction

The computational task of text classification is typically targeting the question of *domain*: Is a text about sports, the economy, local politics, etc. But texts can also be grouped by their *genre*: Is it a business letter, a personal homepage, a cooking recipe, and so on. In this paper, we perform genre classification on newspaper text and are specifically interested in the question whether a text communicates a news report or gives an opinion, i.e., it is an editorial (or some similar opinionated piece). This task is relevant for many information extraction applications based on newspaper text, and it can also be extended from newspapers to other kinds of text, where the distinction ”opinionated or not” is of interest, as in sentiment analysis or argumentation mining.

Our starting point is the work by (Krüger et al., 2017), who presented a news/editorial classifier for English. They demonstrated that using linguistically-motivated features leads to better results than bag-of-words or POS-based models, when it comes to changing the domain of text (which newspaper, which time of origin, which type of content). To transfer the approach to Italian, we assembled a suitable corpus for training and testing, selected preprocessing tools, and adapted the features used by the classifier from Krüger et al. Our results are in same range of the original work, indicating that the problem can be solved for Italian in pretty much the same way. We found some differences in the relative feature strengths, however.

After considering related work in Section 2, we describe our corpus (Section 3) and the classification experiments (Section 4), and then conclude.

2 Related Work

In early work, (Karlgren and Cutting, 1994) ran genre classification experiments on the Brown Corpus and employed the distribution of POS-tags as well as surface-based features such as length of words, sentences and documents, type/token ratio, and the frequency of the words ‘therefore’, ‘I’, ‘me’, ‘it’, ‘that’ and ‘which’. Among the experiments, the classification of ‘press editorial’ yielded 30% errors, and that of ‘press reportage’ 25%. On the same data, (Kessler et al., 1997) used additional lexical features (latinate affixes, date expressions, etc.) and punctuation. The authors reported these accuracies: reportage 83%, editorial 61%, scitech 83%, legal 20%, nonfiction (= other expository writing) 47%, fiction 94%.

The alternative method is to refrain from any linguistic analysis and instead use bag-of-tokens (2003), bag-of-words (Freund et al., 2006), (Finn and Kushmerick, 2003) or bag-of-character-n-gram (Sharoff et al., 2010) models. This has the obvious advantage of knowledge-freeness and yields very good results in the domains of the training data, but, as found for instance by Finn and Kushmerick, a bag-of-words model performs very badly in cross-domain experiments. Likewise, (Petrenz and Webber, 2011) show in their replication experiments that this idea is highly vulnerable to topic/domain shifting: the models largely learn from the content words in the training texts, and these can be very different from day to day, when the news and the opinions on them reflect the current affairs.

(Toprak and Gurevych, 2009) experimented with various lexical features: Word-based features included unigrams, bigrams, variants with surrounding tokens, as well as frequency-amended lemma features (using a tf*idf measure); lexicon features exploited the Subjectivity Clues Lexicon (Wilson et al., 2005), SentiWordnet (Esuli and Sebastiani, 2006), and a list of communication and mental verbs. It turned out that word class features outperform the other classes, with an accuracy of up to 0.857. Specifically, the tf*idf representation was successful. Such frequency-based representations are known to be effective for classical topic categorization tasks, and this study provides an indication that they may also help for related tasks (especially when the class distribution is skewed). Another finding was that plain unigrams beat the larger n-grams and certain context features.

(Cimino et al., 2017) investigated the role of different feature types in the task of Automatic Genre Classification. In this study a set of relevant features is extracted across different linguistic description levels (lexical, morpho-syntactic and syntactic) and a meaningful subset is then selected through an incremental feature selection procedure. The results show that syntactic features are the most effective in order to discriminate between different text genres.

Finally, as mentioned earlier, we build our work on that of (Krüger et al., 2017), who systematically tested a meaningful set of linguistic features. Among several classifiers from the WEKA libraries, the SMO classifiers performed best, and the models based on linguistic features outperformed standard bag-of-lemma approaches across different genres, but the latter still performed very well on the same genre on which they were trained. Krüger et al. then tested which features are most predictive for each class, and related these observations to their original expectations.

3 Dataset

For our study, we built a corpus of about 1000 Italian newspaper articles, which are equally divided into editorials and news articles.

The editorials have been collected from the website of the Italian newspaper *“Il Manifesto”* and we removed headers and footers that serve as metadata for the newspaper, such as “2017 IL NUOVO MANIFESTO SOCIETÀ COOP. EDITRICE”. The news articles are from the Adige corpus[1], a collection of news stories from the local newspaper *L’Adige* categorized into different topics of news, such as *sport*, *finance* or *culture*. The corpus is also annotated with semantic information related to temporal expressions and entities. However, we have not exploited these features since they were not available on the editorials.

Both corpora have been annotated using the *TreeTagger* tool[2] (Schmid, 1994), which provides an annotation of the form WORD, POS-TAG, LEMMA.

In order to reproduce the types of classification features used by (Krüger et al., 2017), some lexical resources are needed. The corresponding Italian vocabulary has been collected from different sources:

- A list of connectives, categorized into temporal, causal, contrastive and expansive connectives, has been obtained from LICO (Feltracco et al., 2016), a lexicon for Italian connectives.

[1] http://ontotext.fbk.eu/icab.html

[2] Future improvements include using a more modern postagger such as UDPipe: https://ufal.mff.cuni.cz/udpipe

	Acc.	Prec.	Recall	F1
L	83,35	86,04	79,42	82,60
P	84,49	85,80	82,50	84,11
U	82,29	80,29	85,38	82,75
L+U	87,75	88,88	86,15	87,50
L+P	87,27	88,46	85,58	87,00
U+P	87,37	87,31	87,31	87,31
L+P+U	89,09	89,64	88,27	88,95

Table 1: Linear SMO results: L: Linguistic features, P: POS tagging, U: Unigrams

	Acc.	Prec.	Recall	F1
L	83,90	84,21	82,75	83,47
P	64,71	63,08	69,49	66,12
U	39,17	43,30	70,00	53,50
L+U	65,00	50,57	73,33	59,86
L+P	72,57	70,37	71,70	71,03
U+P	50,83	50,57	73,33	59,86
L+P+U	61,34	57,83	81,35	67,60

Table 3: Linear SMO results on Amazon reviews and Wikipedia articles

- A list of communication verbs (*say, argue, state*, etc.) has been obtained from the lexical database *MultiWordNet*[3] for a total of 54 entries.
- Sentiment features rely on the *Sentix*[4] lexicon for Italian sentiment analysis, which assigns to each lemma a positive and negative score, plus a score of polarity and intensity.

4 Experiments

Feature	Weight
LING:PRONOUNS	3,5452
LING:TEMPORALCONN	2,0647
LING:SENT_POS	1,8040
LING:NEGATIONS	1,7301
LING:SENT_NEG	1,6609
LING:PAST	1,3686
LING:CONTRASTIVECONN	1,2816
LING:INFINITIVE	1,2230
LING:SENT_ADJ_POL	1,2114
LING:SENT_ADJ_NEG	1,0880
LING:CONDIMP	1,0796
LING:GERUND	1,0653
LING:COMMAS	0,9658
LING:SENT_INT	0,9593
LING:IMPERFECT	0,7801

Table 2: Linguistic features pointing to opinionated text

Feature	Weight
LING:CITATIONS	4,8912
LING:COMPLEXITY	2,6676
LING:PASTPERFECT	2,1070
LING:FUTURE	2,0092
LING:TOKENLENGTH	1,8754
LING:CAUSALCONN	1,7568
LING:SENT_POL	0,9710
LING:VOS	0,7414
LING:IMPERATIVE	0,6871
LING:FSPRONOUNS	0,6518
LING:FPRONOUNS	0,6518
LING:MODALS	0,4237

Table 4: Linguistic features pointing to news text

4.1 Main experiment: feature performance

In our experiments, we were primarily interested in comparing the accuracies obtained by (i) linguistic features, (ii), unigram counts, (iii) part of speech tags counts, and their combinations as indicators for classifying the newspaper articles from the dataset. Four different classifiers from the WEKA library have been tested: linear and polynomial SMO (kernel with $e = 2$), J48 trees and Naive Bayes classifier, with a 10-fold cross-validation evaluation. The SMO classifiers proved to be the most accurate, with the polynomial SMO having marginally higher scores than the linear counterpart. In Table 1 we provide our results obtained with that approach. It can be seen that combining feature sets generally outperforms the individual sets, and in fact the combination of all three yields the best results.

Our set of linguistic features was modeled closely after that of Krüger et al., because we wanted to know how well it can be transferred to languages other than English. These features can be summarized as follows: text statistics (length of a sentence, frequency of digits, etc.); ratio of punctuation symbols; ratio of temporal, causal and other connectives; verb tenses; pronouns (esp. 1^{st} and 2^{nd} person) and sentiment indicators.

[3] http://multiwordnet.fbk.eu/english/home.php

[4] http://valeriobasile.github.io/twita/sentix.html

The set also includes the presence of modal verbs and negation operators, morphological features of the matrix verb (tense, mood), as well as some selected part-of speech and basic text statistic features, as they had already been proposed in the early related work.

The feature weights assigned by the linear classifier are shown in tables 2 and 4 in order to highlight which linguistic features represent good indicators towards one or another type of article, and with how much strength.

The results obtained offer interesting analogies with the English corpus analysed by (Krüger et al., 2017). For instance, pronouns, negations and sentiment represent strong indicators for opinionated texts, while complexity, future, communication verbs, token length and causal connectives are all features pointing towards news reports in both languages. An interesting difference is the role of past tense, which for English had been found to correlate more with news than with editorials, and here it plays a different role.

4.2 Testing domain change robustness

We then evaluated another aspect of the task, viz. domain robustness: we split the news corpus into a training set (categories Attualità, Sport and Economia) and a test set (categories Cultura and Trento) in order to evaluate the robustness of the classifier when unseen categories are submitted. All the classification performances in this setting show a drop of performance of only about 0,03%, demonstrating that the classification performances are not overfitted to the topics of the articles.

Finally, to further test domain change robustness, we tested the classifier – with the model trained on the newspaper corpora – on a set of 60 Amazon reviews versus 60 Wikipedia articles (all randomly chosen). As the results in Table 3 show, the linguistic features perform remarkably robust also on this quite different data. The bad results for unigrams on the one hand are not so surprising, but they have to be taken with a grain of salt, because we employed the same low frequency filtering as in the main experiment: unigrams that occur less than five times are not being considered, in order to reduce the feature space. This might well lead to poorer results for a small data set like the 120 texts used here.

4.3 Replication

Altough we cannot make public all the data we used in this experiment, we uploaded our code on a public repository[5] to provide a description of our implementation.

5 Conclusion

We presented, to our knowledge, the first classifier that is able to distinguish 'news' from 'editorials' in an Italian newspaper corpus. It follows a linguistic feature-oriented approach proposed by (Krüger et al., 2017) for English, who had demonstrated that it outperforms lexical and POS-based models. In our implementation, With an accuracy of 89.09% the distinction between the two subgenres can be drawn quite reliably. Our results are comparable to that of Krüger et al., which indicates (again, to our knowledge for the first time) that their feature space is applicable successfully to languages other than English.

Our central concern for this kind of task is robustness against domain changes of different kinds. To this end, Krüger et al. had worked with different newspaper sources and demonstrated the utility of the feature approach in such settings. While we were not able to assemble large corpora from different papers, we ran other experiments in the same vein, where the first shows that the system is robust against changing the portions of the newspapers (i.e., economy versus local affairs, and so on). In the second one, we applied the classifier, as trained on the newspaper data, to the distinction between Italian Wikipedia articles and Amazon reviews, where the results remained stable as well. We take this as an indication that the classifier captures a general difference between 'opinionated' and 'non-opinionated' text, and not just some 'ad hoc' phenomena of certain newspaper sub-genres.

References

[Cimino et al.2017] Andrea Cimino, Martijn Wieling, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Identifying predictive features for textual genre classification: the key role of syntax. In Roberto Basili, Malvina Nissim, and Giorgio Satta, editors, *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-*

[5] https://bitbucket.org/PietroTotis/classifying-italian-newspaper-text-news-or-editorial/src/master/

13, 2017., volume 2006 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Esuli and Sebastiani2006] Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.

[Feltracco et al.2016] Anna Feltracco, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. 2016. Lico: a lexicon of italian connectives. In *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it)*, Napoli.

[Finn and Kushmerick2003] Aidan Finn and Nicholas Kushmerick. 2003. Learning to classify documents according to genre. *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*.

[Freund et al.2006] Luanne Freund, Charles L. A. Clarke, and Elaine G. Toms. 2006. Towards genre classification for IR in the workplace. In *Proceedings of the 1st international conference on Information interaction in context*, IIiX, pages 30–36, New York, NY, USA. ACM.

[Karlgren and Cutting1994] Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*, COLING '94, pages 1071–1075, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Kessler et al.1997] Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38. Association for Computational Linguistics.

[Krüger et al.2017] Katarina R. Krüger, Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687–707.

[Petrenz and Webber2011] Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393.

[Schmid1994] Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester.

[Sharoff et al.2010] Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web Library of Babel: evaluating genre collections. *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 3063–3070.

[Toprak and Gurevych2009] Cigdem Toprak and Iryna Gurevych. 2009. Document level subjectivity classification experiments in deft'09 challenge. In *Proceedings of the DEFT'09 Text Mining Challenge*, pages 89–97, Paris, France.

[Wilson et al.2005] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP-2005*.

[Yu and Hatzivassiloglou2003] Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.