Computational Style Analysis: Automatic Labelling and Classification of News Leads

Robin Lam, Annie Louis, Kaitlyn Mulcrone, Ani Nenkova, Ethan Roday
Computer and Information Science
University of Pennsylvania

Abstract

Writing style plays a major role in how a piece of writing is perceived by a reader. However, little research has been conducted on developing a method to objectively score writing style. In this pilot exploration into computational style analysis, we propose a method for classifying the opening paragraphs of articles drawn from the New York Times as informative or entertaining and develop a classifier using style and language features.

1 Introduction

With the majority of current events articles, the major priority of the journalist is to inform. In such cases, the journalist may elect to employ a direct writing style, to impart a large amount of information in a concise, easy to digest block of text. Consider for example this lead from a business article:

> *The Russian state oil company Rosneft has acquired $482 million in debt that the troubled oil company Yukos owed to Western banks, Yukos said Wednesday. The step raised the possibility that Rosneft might acquire Yukos's remaining assets. In a statement on its Web site, Yukos said it appeared that Rosneft acquired the debt in December.*

The sentences have basic syntactic structure and the word choice is simple. This lead is designed for high readability for readers who need to stay informed but short on time. In feature writing, however, the journalist is writing as much to entertain as to inform. This gives rise to much more creative expression within an article. The following lead, also from the business section, illustrates an effective use of literary techniques.

> *Congratulations! You have just been named chief executive. Here's a bit of advice: don't unpack your bags. Not only is the imperial C.E.O. a thing of the past, so, too, is the idea of the long-tenured chief executive, Strategy & Business says in its summer issue.*

The sentences in this lead have much more dynamic lengths. The various punctuation marks force the reader to pause while reading and the author addresses the reader explicitly. This lead is informative, but the way in which it was written also makes it entertaining to read. Any reader familiar with news publications can easily distinguish the differences between these two approaches, but can a machine be taught to label informative or entertaining articles with similar success? In this paper, we elaborate on the steps taken to answer this question:

- We compile a corpus of news articles to use as a basis for classifier development (Section 3).
- We propose a method of automatically labelling news leads as written creatively or informatively through the use of topic word densities, topic word coverages, and news abstract comparisons (Section 4).
- We develop a feature set based off lexicons, syntax, and semantics to be used in supervised machine classification (Section 5).
- We demonstrate the validity of the developed writing style classifier through manual verification (Section 6).

2 Motivation

Writing style is a very broad and often vague term used to describe the way in which an author writes to his or her audience. Thus far, there has been relatively little research done on writing style analysis, but the applications of a writing style classifier are numerous. Stylistic writing can be difficult for non-native speakers to interpret and even more difficult for machines to translate. With a writing style classifier, one would be able to use the classifier scores to identify sentences that are difficult to understand or translate. What is more, the classifier can be adpated as a filter for online searches to only yield texts with a given style. In the classroom setting, students often sacrifice their individual writing styles for an educator's approval, but many argue that learning to

write well should depend on more than an occasionally arbitrary opinion. The development of a writing style classifier would put an end to such practices by providing a objective writer feedback.

3 Corpus Development

The emphasis journalists place on writing quality leads make the first few sentences of news articles an excellent indicator of the writing style of the entire piece. We assembled a corpus comprised of 30,619 articles from *The New York Times* in 2005 and 2006. Furthermore, we sorted the articles in our corpus into four genres, as depicted in Figure 1, using topic tags given by the *The New York Times*.

Genres	Number of Articles
Business	13,319
Science	2,938
Sports	11,528
Politics	2,834
Total	30,619

Figure 1: Number of articles per genre in our corpus

We gave consideration to the genre of articles throughout our analysis because the proportion of articles written solely to inform and those written to inform and entertain varies by genre. For example, in the business genre, one would expect there to be a larger number of purely informative articles, whereas in the sports genre, one would expect that opposite. Also using labels given by *The New York Times*, we were able to identify and extract the leads from each article.

To prepare the articles, we ran Stanford Core NLP (Toutanova, Klein, Manning, Singer 2003) to parse the articles. The program helped us extract sentence breakdowns, words, lemmas, part of speech labels, and dependencies that are used in automatic labelling and/or classification.

4 Automatic Labelling of Leads

4.1 Developing Statistics for Automatic Labelling

Journalistic conventions dictate that news leads for articles written purely to inform should contain all the essential information of the article. In contrast, articles also emphasizing entertainment value use leads to draw readers in, often employing literary techniques such as funnels or anecdotes. With this observation in mind, we claim that topic word (TW) analyses of leads written in one of these styles will be significantly different from leads written in the other style. Topic words were identified for each article using a topic word tool developed by Louis and Nenkova (2012). We calculated two topic word statistics, density and coverage, as defined in Figure 2, for each lead.

$$\text{TW density} = \frac{\text{\# of topic words in lead}}{\text{\# of words in lead}}$$

$$\text{TW coverage} = \frac{\text{\# of unique topic words in lead}}{\text{\# of unique topic words in article}}$$

Figure 2: Topic word statistic definitions

Note that both topic word statistics generate values ranging from 0 to 1. A TW score, the sum of a lead's TW density and TW coverage, was assigned to each lead. We hypothesize that leads from articles written only to inform have larger topic word densities and larger topic word coverages than leads from articles written also to entertain.

Another statistic, independent of topic words, was developed to aid in automatic labelling based on newspaper abstracts. Abstracts were not available for all articles, but, for articles that had abstracts of at least 25 words, a ratio was calculated. First, for each word in the lead, a tuple was created containing the word and it's part of speech. The same was done for each word in the abstract. The ratio used for automatic labelling is the percentage of word, part of speech tuples in the abstract that are also in the lead. Because newspaper abstracts essentially serve as summaries of articles, they often use direct writing styles and therefore lack creative uses of language. We claim that articles written with a direct style will have more words in common with the abstracts and therefore have higher abstract ratios.

Using our topic word scores and abstract ratios, we developed two independent methods of automatically labelling news leads.

4.2 Evaluating the Automatic Labels

4.2.1 Topic Word Score Evaluation

To evaluate the accuracy of the labelling method based on topic words, we first needed to manually label leads to use for comparison. A selection process was developed to eliminate the process of reading and labeling leads by hand. The selection method for leads to be manually labelled is best understood using an example, so consider all the leads in the business genre. First, the leads were sorted from lowest TW score to highest. By our claim, this list is also supposed to be sorted from most entertaining to most informative. Second, the cut-off for the top five percent of leads in this list was calculated. Third, twenty of the one hundred leads occuring just before the five percent cut-off were randomly selected. These twenty leads were read and manually labelled on a binary scale by one member of our team. We recorded the average topic

word density and topic word coverage for these twenty leads. Continuing our example, let's say that 17 of these leads contained a creative writing style and 3 contained a direct writing style. This gives the classifier an 85 percent accuracy rating because the top five percent were supposed to be entertaining. This selection process was then repeated again, but with a cut-off percentaged increased by five percent. The cut-offs used were 5, 10, 20, 25, 30, 35, and 40 percents. This process of selecting twenty leads and recording the mean topic word density and mean topic word coverage was repeated until the classifier's accuracy fell below 70 percent. When the accuracy reached below 70 percent, the selection process was repeated starting from the bottom five percent and, again, progressed through the cut-offs until the accuracy fell below 70 percent. In this manner, the classifier's accuracy has been assessed on business leads. In our research, we repeated this process once per genre and the results of the assessment can be found in Figure 3.

Genre	Cut-Off	Accuracy	Density	Coverage
Business	5%	85%	4.1255	12.06939
	10%	75%	6.3541	16.93903
	15%	70%	6.7124	21.76002
	-5%	80%	15.5767	80.1984
	-10%	95%	14.7913	70.7823
	-20%	85%	14.6077	59.1748
	-25%	80%	13.4741	56.2021
	-30%	75%	13.3668	52.1775
	-35%	55%	11.6314	50.5699
Science	5%	95%	3.4842	6.7414
	10%	80%	5.0273	10.0194
	20%	90%	7.0790	14.7521
	25%	75%	7.5180	16.7658
	30%	55%	8.7371	17.8752
	-5%	80%	15.6205	52.1873
	-10%	90%	13.3936	47.2335
	-20%	65%	13.0102	39.3692
Sports	5%	70%	4.5368	13.9893
	-5%	95%	12.8915	73.9424
	-10%	65%	12.6522	63.6156
Politics	5%	70%	3.7017	13.8463
	-5%	95%	10.1691	87.8255
	-10%	100%	12.0677	73.7242
	-20%	90%	14.4282	59.8409
	-25%	85%	11.4911	58.8002
	-30%	85%	11.5084	55.7594
	-35%	85%	10.1669	53.9689
	-40%	60%	12.0707	48.9722

Figure 3: Topic word classifier accuracies with average density and coverages

4.2.2 Abstract Ration Evaluation

A similar but slightly different approach was used to evaluate the accuracy of abstract ratios. For each genre, an average abstract ratio was computed and all leads below the average were considered to be entertaining, while all leads above the average were informative. Next, the bottom ten percent of the entertaining leads and the top ten percent of informative leads for a given genre were selected. Note that these two groups of leads are not necessarily the same size. Using the same random sampling technique employed for topic word evaluation, we evaluated twenty of the leads in each of these ten percent groupings and recorded their accuracies. Leads were evaluated at ten percent intervals until the accuracy rate dropped below seventy percent. In one of the "halves," we labelled all leads occurring at or prior to the seventy percent accuracy cut-off to be entertaining and, in the other "half", we labelled all leads occurring at or prior to the seventy percent cut-off to be informative.

Genre	Cut-Off	Accuracy	Abstract Ratio
Business	0-10%	80%	8.78
	10-20%	65%	16.14
	20-30%	75%	20.21
	30-40%	70%	23.81
	50-60%	70%	63.91
	60-70%	80%	68.24
	70-80%	80%	73.57
	80-90%	80%	79.37
	90-100%	90%	85.36
Science	0-10%	90%	5.23
	10-20%	90%	9.26
	20-30%	85%	11.71
	30-40%	80%	13.67
	40-50%	85%	15.78
	90-100%	50%	72.05
Sports	0-10%	80%	8.90
	10-20%	85%	14.70
	20-30%	75%	17.91
	30-40%	80%	20.97
	80-90%	65%	62.69
	90-100%	90%	72.54
Politics	0-10%	95%	9.62
	10-20%	75%	17.01
	20-30%	50%	22.28
	50-60%	65%	58.91
	60-70%	80%	61.84
	70-80%	70%	66.38
	80-90%	75%	72.62
	90-100%	90%	81.05

Figure 4: Abstract classifier accuracies with average abstract ratios

5 Feature Set Development

Before performing supervised machine training, we needed to develop a feature set. Our feature set includes over one hundred features based off lexicons, syntax, and

semantics. The following paragraphs detail how each feature was computed and why we chose to use it as a feature.

5.1 Lexical Features

Word Length We developed two features based on word length: average word length and word length variance. To distinguish words from symbols, contractions, and punctuation, tokens in a given lead were filtered by their part of speech tags. Refer to Figure 4 for a list of all tags and parts of speech that were considered to label words. We computed the average word length by sentence and averaged those averages for the lead average. Word length variance was the variance of the average sentence word lengths. We chose to use average word length as a feature because word lengths are often associated how difficult a word is to learn. Furthermore, word length variance is one way to measure how dynamically written a lead is.

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition/subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
PRP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund/present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person sing. present
VBZ	Verb, 3rd person sing. present
WDT	*wh*-determiner
WP	*wh*-pronoun
WP$	Possessive *wh*-pronoun
WRB	*wh*-adverb

Figure 5: The tags and the parts of speech they represent that were considered to label words. Parts of speech that did not indentify words were possessive endings, symbols, and all punctuation.

Part of Speech Composition We also calculated the densities of certain parts of speech within sentences. These parts of speech included *wh*-determiners, *wh*-pronouns, *wh*-adverbs, adjectives, adverbs, proper nouns, person pronouns, and verbs. For adjective and adverb densities, both numbers can be interpretted as a measure of how descriptive an author is. On the other hand, many have claimed that good writing limits the usage of adjectives and adverbs and instead relies on verbs to drive sentences. In an online article written to advise high school students on writing news articles, *The New York Times* explicity states, "Don't clutter up the lead, or the article, with adjectives and adverbs...writing that employs vivid verbs and telling details is much more powrful than writing that leans heavily on modifiers."

Present and Past Verbs A present tense verb ratio was calculated by counting the number of present tense verbs in a given lead and dividing by the total number of verbs in the lead. We identified present tense verbs by the parts of speech tags VBG, VBP, VBZ, and VB and performed a similar calculation to find the past tense verb ratio, using the tags VBD and VBN (please refer to Figure 4 for the abbreviation meanings). The present tense verb ratio is useful because newspaper articles are traditionally written in past tense. We chose to also measure the present tense verb ratio because present tense verbs would provide nuance and present tense writing is often found to be more engaging than past tense writing.

Passive Voice We detected passive voice in leads by searching through the dependencies returned by Stanford Core NLP for the phrase *nsubjpass*. *Nsubjpass* is a label used to identify noun phrases which are the syntactic subject of a passive clause. A lead either recieved a score of 0 or 1 for the passive voice detection depending on whether or not it contained a poassive nominal subject. We have two reasons for using passive voice as a feature. First, writers are typically told to avoid using passive voice because it is not as forceful as active voice. Second, we observed while reading samples from our corpus that articles including passive voice generally discussed fatalities and these articles tended not use large amounts of creative language.

Modifier Density Using the labels found in the lead dependencies, we counted the number of adverbial clause, adverbial, adjectival, appositional, infinitival, noun compound, noun phrase as adverbial, numeric, participial, prepositional, prepositional clausal, quantifier phrase, relative clause, and temporal modifiers. This count was then divided by the number of words in the lead to yield the lead's modifier density. This statistic is similar to the adverb and adjective densities but provides a more comprehensive statistic for measuring the use of descriptions.

Verb, Adverb, and Adjective Novelty We wanted to find a way to measure the novelty of certain parts of speech within a given lead. To accomplish this, word frequencies were gathered for every word used in the leads of a given genre. An individual word's novelty score was

simply its frequency. From these word scores, we calculated lead novelty scores, lead novelty variances, and lead novelty means. A lead's verb novelty score was the sum of the novelty scores for all the verbs in the lead. A leads novelty variance was the variance of novelty scores of the words of a particular part of speech and the novelty mean was the average of the novelty scores of the words of a particular part of speech. We predict that leads with higher verb, adverb, and adjective scores have more style.

5.2 Syntactic Features

Sentence Length Informative leads are expected to contain a vast amount of information, but, with the length restrictions placed on leads, we observed that this often led to long sentences riddled with details. We counted the length of sentences within a lead by the tokens labelled by Stanford Core NLP and kept their average as a feature. We predicted that leads with shorter sentences have more writing style while leads with longer sentences have less. Literature on writing often advises aspiring authors to vary the lengths of sentences as a way of keeping their audience engaged. For this reason, we included a sentence length variance feature and predicted that sentences with higher levels of writing style also have higher sentence variances. These variances were calculated by lead and excluded punctuation.

Sum of Sentence and Word Length Variances Based off our predictions on the individual variances, we wanted to see if there was a correlation between the two statistics. We predicted that the sum of sentence and word length variances would be higher with stylistically creative leads.

Topic Word Density A feature piece in a newspaper is, by definition, never written to be just informative. The online book, *Campus Weblines* published by *The New York Times* points out "feature articles usually begin with a delayed lead - an anecdotal or descriptive lead," a technique that we also noticed in the leads we read. These delayed leads leave their most topically relevant details until the end of the lead, leaving the first few sentences to capture the attention of the audience. This information led us to investigate topic word density variances between sentences in a lead. Additionally, we computed the difference between the topic word densities of certain parts of a lead: first sentence versus others, last sentence versus others, and first sentence versus last sentence. Ideally, the leads with high first-other differences would use a direct writing style, while leads with high last-other differences would use a creative writing style.

Punctuation We also explored punctuation densities and coverages for leads in our corpus. The types of punctuation we counted are periods, question marks, exclamation points, commas, colons, semi-colons, parentheses, apostrophes, quotation marks, hyphens, rectangular braces, and grave accents. Punctuation density was calculated by counting the punctuation marks found in a lead and dividing that number with lead's word count. Coverage was calculated using the same method topic word coverages were done, by dividing the number of unique punctuation marks in the lead by the number of unique punctuation marks we counted.

Prepositions The use of prepositions in a sentence make a sentence more complex because they allow writers to include more information in that sentence. Given that many informative leads are written by including as many important details as possible in as few sentences as possible, we predict that informative leads will have a larger number of prepositions. For each lead, we counted the number of prepositions by sentence and divided those counts by the lengths of the sentences. The average of these counts became the lead's prepositional mean. Additionally, we included in the feature set the variance of the prepositional densities by sentence and the range of the densities.

Sentence Specificity Using the classifier developed by Louis and Nenkova, we were able to classify sentences as general or specific. A lead's specificity score was the average of the specificity scores of its sentences. Additionally, we created another feature by averaging the probabilities of a sentence falling under a specific class rather than the binary labels. We predicted that leads with direct writing styles would be rated as specific more frequently than those with creative writing styles given the nature of informative writing.

Modifier Distances *Campus Weblines* also advises aspiring writers that the distance between modifiers and what they modify should be as minimal as possible. For every modifier found (see Modifier Densities feature), we used sentence dependencies from Stanford Core NLP to calculate the average distance between the governors and the dependents for all the modifiers in a lead. The smaller the distance between the governor and the dependent, the less cognitive processing it takes a reader to understand the sentence. Therefore, we claim that sentences with direct writing styles will have smaller modifier distances.

Noun, Verb, and Prepositional Phrase Lengths Earlier work done by Chae and Nenkova (2009) motivated our inclusion of phrase feature statistics. We computed the average lengths of noun, verb, and prepositional phrases in a lead as well as the average number of phrases for each part of speech.

5.3 Semantic Features (Sentiment Dictionaries)

We included in our feature set many features derived from sentiment dictionaries to help us measure meaning in the leads.

MRC Database Gilhooly and Logie (1980) compiled a list of 1944 words, each with age of acquisition, imagery, concreteness, familiarity, and ambiguity measures. If a

given lead contained one or more of the words from the MRC database, we computed mean, median, and variance scores for each of the five measures for the lead. If a given lead did not contain any MRC words, then it's MRC database scores were zeroes. We predict that leads with more creative style will also have higher levels of imagery than those written with a direct style.

MPQA Database The MPQA Database (Weibe and Hoffmann 2005) we were able to apply by finding all the words in a given lead that were also in the database. For each lead, we averaged the subjectivity and polarity scores for the MPQA words and assigned those averages to be features.

Affective Text Database Strapparava and Valitutti (2004) developed lists of words that capture anger, disgust, fear, joy, sadness, surprise. We created six word density features from these emotions and predicted that the positive emotions, joy and surprise, would have higher word densities in the creatively written leads.

Regressive Imagery Dictionary From the Regressive Imagery Dictionary (Martindale, 1975, 1990), we measure the density of words associated with primary and secondary thought processes, and emotion words. In addition to those three features, we calculated a fourth that was the difference between the density of primary thought words and secondary thought words. The dictionary itself is sorted into 29 categories of primary process thought, 7 categories of secondary process thought, and 7 emotions. We created separate word-density features out of these 43 mini-dictionaries.

Financial Dictionary Loughran and McDonald (2011) developed severals word lists divided into the categories: negative words, positive words, uncertainty words, litigious words, strong modal words, and weak modal words. We computed individual word density scores for each of the word lists and, in this way, developed six features from the financial word dictionary. We hoped this dictionary would help us capture the subtleties of the business and some of the political articles in our corpus.

5.4 Feature Set Analysis

For analysis, our leads were sorted into smaller groupings based off our initial two: the automatically classified and the manually classified. Our set of automatically labelled leads was the union of leads labelled with the topic word classifier and the abstract ratio classifier. The set of manually labelled leads were the leads we hand labelled while verifying the accuracies of the classifiers. Of the automatically classified leads, we randomly separated the leads into a training set and a testing set such that the training set was ten times larger than the testing set. We created a smaller test set, called "overlap," from our hand classified leads. This small set of twenty-nine is the collection of leads that were read by both of our manual labelers and given the same judgement.

To evaluate our features, we ran T-Tests based on the score averages of the manually labelled leads. We chose the manually labelled instead of the automatic because we wanted to maximize the classifier's accuracy when analyzing the manually labelled leads. All features that had a p-value of under 0.05 were kept, while the others were excluded.

Set	Size	Baseline	Accuracy
Automatic	1813	61.72%	79.7022%
Manual	1319	55.50%	74.4503%
Overlap	29	62.07%	93.1034%

Figure 6: Classifier accuracies when applied to automatically classified and manually classified leads.

References

A. Louis and A. Nenkova. 2011. *Automatic identification of general and specific sentences by leveraging discourse annotations.* In *Proceedings of IJCNLP.*

J. Chae, A. Nenkova. 2009. *Predicting the Fluency of Text with Shallow Structural Features: Case Studies of Machine Translation and Human-Written Text.* In *Proceedings of EACL*: 139-147.

C. Martindale. 1975. *Romantic progression: The psychology of literary history.* Washington, D.C.: Hemisphere.

C. Martindale. 1990. *The clockwork muse: The predictability of artistic change.* New York: Basic Books.

T. Loughran and B. McDonald. 2011. *When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.* In *Journal of Finance*: 35-65.

C. Strapparava and A. Valitutti. 2004. *WordNet-Affect: an affective extension of WordNet.* In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*: 1083-1086.

M.D. Wilson. 1988. *The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2.* In *Behavioural Research Methods, Instruments and Computers, 20(1), 6-11.*

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.* In *Proceedings of HLT-EMNLP.*

C. Chang and C. Lin. 2011. *LIBSVM: A library for support vector machines.* In ACM Transactions on Intelligent Systems and Technology vol. 3 issue 3: 1-27.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.* In *Proceedings of HLT-NAACL 2003*: 252-259.

A. Louis and A. Nenkova. 2012. *Automatically assessing machine summary content without a gold-standard.* To appear in *Computational Linguistics, 2012.*