Computing and Informatics, Vol. 35, 2016, 1027–1049

BLOG STYLE CLASSIFICATION: REFINING AFFECTIVE BLOGS

Martin Virik, Marian Simko, Maria Bielikova

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
e-mail: {martin.virik, marian.simko, maria.bielikova}@stuba.sk

Abstract. In the constantly growing blogosphere with no restrictions on form or topic, a number of writing styles and genres have emerged. Recognition and classification of these styles has become significant for information processing with an aim to improve blog search or sentiment mining. One of the main issues in this field is detection of informative and affective articles. However, such differentiation does not suffice today. In this paper we extend the differentiation and suggest a fine-grained set of subcategories for affective articles. We propose and evaluate a classification method employing novel lexical, morphological, lightweight syntactic and structural features of written text. The results show that our method outperforms the existing approaches.

Keywords: Blog, information processing, writing style, web genre, classification, natural language processing

1 INTRODUCTION

In line with adoption of read/write web vision [20] to praxis, new channels of communication and information interchange have emerged on the Web. They have shifted the notion of web users from passive visitors into active content creators, who produce information on their own instead of consuming it. Information interchange is facilitated through new social media such as blogs, microblogs or general-purpose social networks.

In the current Web, after two decades of evolution, blogs represent a genre that stands between static information pages and live forums [18, 11]. They have become

a tool for ordinary users to share information, ideas or even emotions, creating a heterogeneous mass of user-generated content filled with unique information related to individual as well as society-wide issues. Since the blog articles are typically weakly structured, the extraction of valuable information is becoming increasingly difficult to handle.

Even though there are no restrictions or limits to content or form in blogs, users seem to create writing styles and genres spontaneously reflecting their intention and current emotional state. Information processing in such content should consider writing style specifics. There is more than 2.5 million of blog posts written every day [38]. A certain degree of guidance should be given to web users to overcome "the fuzziness of the blogosphere" [11]. Modern search and recommendation services should be aware of genres and consider them as additional criteria when compared with filtering conventional web content [3]. For example, by using genre information, it would be possible to filter articles with low information value (e.g., personal diaries, in contrast with technical blogs), enabling search engines to focus on relevant sources only, or, alternatively, to recognize sentiment about a specified object. Bringing order into the blog genres and writing styles would help readers to find their preferred blog and it would help writers to establish a community of readers.

Two most common writing styles recognized in blogs are informative and affective [23]. *Informative* blogs are focused on delivering information with only a small or no amount of personal influence. They usually include tutorials, news reports, recipes, etc. On the other hand, *affective* blogs aim at expressing personal opinions or emotions over a particular subject. Examples include comments, diaries or stories.

Having the ability to filter affective articles would be helpful in situations, when readers are searching for actual news stories and are not interested in out-of-date personal commentaries. On the other hand, readers often look for stories experienced by bloggers on specific location (e.g. stories about vacation in Slovakia). Here, it would be helpful to extract affective articles and to decide afterwards whether it is a story or something else. Another scenario represents a situation when a reader is looking for some kind of technical tutorial or an advice (e.g. a function usage in Microsoft Excel or replacing a pollen filter in Ford Fusion). The full-text search would return a lot of articles, wherein the author tells only a story of an attempt to complete the task (in many cases an unsuccessful one). In order to filter out this kind of articles, we would need to differentiate between informative and narrative, or even emotional writing styles. To achieve these goals, we need to employ an approach that would classify blogs and articles into a proper space of well-defined genres and writing styles.

In our work we follow informative and affective styles classification and we suggest a set of subcategories for affective articles that make a dominant category of affective blog style fine-grained (Section 3). By identifying and combining two dichotomies of further differentiation (writers' intention and sentiment), we obtain new blog categories: reaction to an event, rational reflection, reaction to a personal experience and rational story. We propose novel features for blog classification ac-

cording to this differentiation (Section 4). Our contribution is that we consider not only word usage, lexical and morphological attributes (typical for state-of-the-art approaches), but also more complex features such as sentence syntax or text structure. The conducted experiments confirm that utilization of such features leads to reasonable results (Section 5). The experiments were performed on blogs in not well researched Slovak language consisting of real world posts from the most popular Slovak blogging platform with baseline annotations provided by multiple real web users. The obtained experimental results are very promising and they outperform closely related baseline method in informative and affective blogs classification.

Overall, our contributions to the state of the art can be summarized as follows:

- extended classification space aimed at further differentiation of predominant affective blog articles,
- novel linguistic features for classification, not included in blog classification before,
- evaluation of our approach utilizing real world blogs compared with annotations provided by multiple real users, and
- evaluation on the Slovak language that has not been excessively involved in similar research before; with obtained observations applicable to a bigger part of the generic language family.

2 RELATED WORK

2.1 Blogs in the Context of Web Genres

From the beginnings of the Web, researchers have been concerned with the genres that emerged especially from the user generated content. Blog became a subject of web content classification as it became one of the seven identified main genres of the Web [28, 22].

Multiple works deal with automatic genre identification for web pages in line with Santini's classification [28]. Mehler et al. contributed with a collection of approaches and methods of automated web genres recognition and computational model creation, building a solid foundation for genre and writing style research on the Web [22]. Even more recently, the task of web genre classification still attracts attention of researchers who continue to pursue improvement in automatic web page classification [26, 36].

Alongside studies on web genres, blog itself became a subject of research and further analysis. Various works study blog as a novel genre emerged along the expansion of the Web, analyse it from social, psychological, cultural and linguistic perspectives and provide own taxonomy of subgenres [4, 11]. Blood identifies three types of blogs based on writer's intention or the functionality of the text [4]:

- filter – contains mostly links for other potentially interesting pages with additional user notes. The function of this type is to inform,

- personal diary – the content is focused on the author's personal topics, thoughts and feelings. The function of this type is to present opinion or emotion,
- notebook – a combination of both previous types.

Blood in her study also mentions the dominance of personal diaries over others in terms of quantity. On the other hand, the results of several empirical studies point to the dominance of filters in the matter of popularity. Krishnamurthy arguments that the most visited and most commented blogs are those with the most objective information [17]. He proposes a space of four classes defined by a combination of two dichotomies: personal vs. topical and individual vs. community.

Hoffman studies blog evolution and recognizes four blog genres according to the purpose of blog: diary blog, friendship blog, career blog and commercial blog, and identifies subsequent blog sub-genres (such as private journals, travel blogs, fiction blogs; topic blogs, techblogs, political blogs, artblogs; journalistic blogs, edublogs, medblogs; and corporate blogs, company blogs; respectively) [11].

Ni et al. differentiate between affective and informative blogs [23]. The first type covers articles about the webloggers' personal feelings, thoughts or emotions. The second type is about technologies and various kinds of informative news.

2.2 Blog Style Classification

Although the research in the domain of blogs has been very active, when considering blog classification in recent years, it focuses rather on blog topic classification [32, 19, 13, 35, 2, 6, 31, 37]. Only few works deal with blog genres classification [8, 23, 27].

Elgersma and Rijke follow up the vertical dimension of blog classification and automatically differentiate between personal and non-personal blogs [8]. While considering a set of features (term frequency, usage of pronoun, inner and outer links and hosting domain name), they built five machine learning classifiers. For their evaluation they used a hand labeled set of 152 blogs (76 personal and 76 non-personal) randomly sampled from a blog collection. Using a decision table classifier, they achieved 90.13 % accuracy. However, this result was obtained mostly due to features, which are applicable only to specific datasets. For example, domain name cannot be used as a feature in datasets, where all articles are from the same domain.

Similar approach is presented by Ni et al. [23]. Their goal was to create a classifier for distinguishing between informative and affective blogs. Informative blogs include news articles, technical descriptions, general knowledge and objective commentary on world events. On the other hand, affective blogs are mainly represented by personal diaries. The approach of Ni et al. is based on term frequency using tf-idf score and a machine learning classifier. For evaluation they gathered a set of more than 7 000 labeled Chinese blogs and they managed to achieve 92 % accuracy. Applying the classifier to nearly 100 000 blogs confirmed the dominance of the affective blogs.

Some researchers aim to classify blogs to hybrid or mixed topic-genre categories, where some classes have notion of writing style, or consider blog as separate category

among other genres. For example, Qu et al. categorized blogs four groups: personal diary, politics, news and sports [27]. By employing simple unigram-based feature with Naive Bayes classifier, they obtained reasonably good classification accuracy of 84 %. The experiment however employed together only 120 blogs.

Although the blog genre classification yields reasonable results for high-level genre differentiation, there is absence of further lower-level genre analysis. Prior research has shown that approximately 85 % of blogs are affective [23]. Such amount of blogs should be further categorized reflecting emotions with various distributions at a different level of granularity. However, many of affective blogs do not necessarily have a distinct emotion (positive or negative), but still differ in writing style. It is important to seek for an alternative distribution of blog styles in order to find additional general yet natural categories not based on emotions only. It is also important to extend classification space while preserving or even further improving classification accuracy. In comparison with blog topic classification, not much research work has been devoted to blog genre classification in recent years; despite that the problem of information overload becomes even more important and need of fine-grained filtering of blog will be increasingly important with growing numbers of blog posts every day [38]. In this work, we aim to fill this gap by extending blog genres classification space.

2.3 Perspectives of Blog Classification

A potential basis for accuracy improvement lies in the used feature selection of the classification methods. Recent advances are reported in both general purpose text classification [1, 41, 12, 30, 15] and sentiment analysis [24, 39, 33] that we consider to be related with the genre identification. However, to the best of our knowledge, a comprehensive analysis of blog content has not been performed extensively before. The same applies to text classification by style, wherein the lack of research in contrast with the traditional topic-based classification is also reported by other researchers [30]. The aforementioned approaches consider only lexical or morphological features and they use large training sets or additional non-universal features (e.g., domain name), which are often not available in or not applicable to other domains, to achieve high accuracy. Writing style classification requires the analysis of deeper linguistic features that can further contribute to the more extensive distribution of blog styles.

Non-marginal implications for blog classification may originate in microblog research. Due to the high popularity of microblogging platforms, a substantial amount of research has been devoted to the phenomena in terms of both topic classification (e.g., [21]) and sentiment analysis (e.g., [16]). Specificities of microblogs, both linguistic and structural, however, make microblog classification a bit distant from traditional middle-length text classification.

Another important aspect in blog classification is language of blog posts. Not surprisingly, analysis of English blogs is prevalent. Recently, however, we witness an increase of works related to other languages, e.g. Chinese or Romanian [31, 37].

We follow this trend in our work as we deal with Slovak blogs classification that has not been researched yet. Furthermore, our findings can be generalized and applied also to other languages where inflection is present to a non-trivial extent.

To fill the gap in the blog style classification, which currently does not reflect the increasing need for fine-grained differentiation of blog styles, we follow the differentiation between informative and affective styles and propose new classification for affective blogs. Compared to alternatives (wherein differentiation is based on a dominant emotion type), the advantage is that this approach considers the fact that many blog posts do not have a specific dominant emotion and are rather neutral or rational. Attempts to identify emotion in such posts lead to inaccuracy. The other advantage of the proposed differentiation is that it brings a second dimension, which makes it possible to distinguish between stories and opinions. We adapt the classification methods to further classify affective articles into new classes while utilizing a comprehensive linguistic analysis of underlying blog corpora. To the best of our knowledge, no such advanced analysis of blog content has been tackled so far.

3 EXTENDED CLASSIFICATION SPACE

As a response to calls for providing more guidance to web users [11, 8], we follow the differentiations present in aforementioned research streams and we further extend classification of affective blogs, which represent the majority of blogs [23]. Our aim is to provide fine-grained classes that characterize affective blogs to help recognize writing style, and, as a result, to facilitate fulfilling information needs of web users. Based on empirical analysis of a large number of blog posts, we differentiate affective blogs from two perspectives that we identified as important.

We have observed that two forms of information presentation are prevalent. They are related to how the author intends to present information he wants to share on blog. The blogs can be divided into the two groups:

- reflective – the author's intention is to express his opinion on a particular subject. He/she usually analyses his/her experience explicitly. The blog article usually covers a situation ongoing in the present and it has consequences in the future.
- narrative – the author's intention is to tell a story that happened in the past. It is rather a logical recount of the experienced event.

Note that reflective articles can have narrative sections, but they are still reflective in most sections and vice-versa.

After the empirical analysis, we found another aspect – an intensity of emotional presentation – to be distinctive. A center of attention of the second group classes is author's sentiment and emotional state during writing. It is often related to the length of interval between the time of the experience and the time of writing. We recognize the following two groups:

- emotional – an article is author's immediate reaction to a particular event. He/she needs to express his/her opinion or tell a story as fast as possible and does not think about a word selection or a sentence structure.
- rational – the author thinks about his/her article before writing to produce a rather rational message. The word usage and the text structure are much more consistent over the whole article.

By combining these two groups we cover various types of blog posts. First, we can accommodate a direct reaction to an event or a situation. Bloggers often write a short personal comment on a non-personal (social, political, etc.) event they have read or heard only shortly before writing the comment. For example, reactions on the results of the elections in a parliament or a football match. Secondly, we cover also a rational reflective article on a subject that the writer deals with over a period of time. The article has been reviewed and rewritten usually several times before publishing and therefore has much more balanced structure and arguments. Thirdly, our classification space accommodates direct reaction to something personal that happened to the writer shortly before writing (usually on the same day). For example, the blogger writes a short log, where he/she expresses his/her current thoughts and feelings about it. Finally, we cover also rational story from the writer's past (i.e., vacation, a tale from childhood). For example, the author writes about a personal story that usually spreads over more than one day. The text is often supported by selected pictures.

The proposed categorization is a result of a need to differentiate between reflective, narrative, emotional and rational blogs when processing information from blogs, e.g., searching in the blog space, filtering relevant blogs or simply adapting access to blogosphere to reflect particular blog-related preferences. It is important to note that some blog posts may, however, show an equal proportion of attributes from both class groups (especially the reflective/narrative one), which can be a disadvantage in classification.

The importance of the proposed differentiation is in the focus on the author, his/her intentions and emotional state. We believe that it better reflects the distribution of blogs in the growing blogosphere. The main aim of our work is classifying blog articles in the proposed classification space.

4 BLOG FEATURES EXTRACTION

There are multiple factors that affect the writing style. They include specific vocabulary and textual, lexical, syntactical and other linguistic features [14].

In our work we adopt feature analysis for Slovak language. However, our approach may be applied to a bigger group of inflected languages. The strong morphological variability of inflected languages offers more possibilities in word usage. However, it also brings a number of issues in computational processing (e.g., difficult lemmatization, morphological tagging). The particular advantage of such languages

is the possibility to hide subject in a sentence. Usage of subjects (especially personal pronouns) is a valuable attribute of the author's intentions and his emotional state. Thanks to the inflection, predicates have the ability to reflect some of the grammatical categories of the hidden subjects, giving us the possibility to analyse them even though they do not explicitly appear in the text.

Note that we do not employ lexicon-based approach in this paper, typically employed for the task of sentiment analysis. Our main aim is to explore multi-level structural properties of text, with a full potential of automatic feature extraction (in contrast with lexicon-based approach, wherein a language-dependent dictionary of emotive words is necessary).

Though aiming at inflected language, our approach can be also used for different language groups wherein more comprehensive linguistic features are present (as some features of inflected language are to a certain extent present in many non-inflected languages).

For blog features extraction we propose a text pre-processing method based on four levels of linguistic processing:

1. lexical analysis,
2. morphological analysis,
3. lightweight syntactic analysis,
4. structural analysis.

4.1 Lexical and Morphological Analysis

The aim of lexical and morphological analysis is to examine the usage of words and word classes to capture author's vocabulary. The first step of our method is tokenization, where we identify four types of tokens: words, numbers, punctuation and separators. Afterwards we assign lemmas and perform morphological analysis of words (note that this step may differ with regards to a particular language used).

The result of lexico-morphological analysis is a collection of six lexical and eight morphological features (see Table 1 and Table 2).

4.2 Syntactic Analysis

In this step we perform lightweight syntactic analysis by processing sequence of morphologically annotated tokens from previous step to capture sentence structure and relations between sentences, words and other tokens. The result of syntactic analysis is a collection of eight features (see Table 3).

Predicate candidates are the verbs that we have used to identify the sentences in the previous step. In our feature set we investigate their dominant tense, person and number categories along with the intensity of this dominance.

ID	Name	Description
Lex_1	Special characters frequency	frequency of the following characters: @, #, $, %, , &, , -, , =, +, ¿, ¡, [,], /, \, |
Lex_2	Word count	number of alphanumeric tokens
Lex_3	Unique lemma count	number of unique identified lemmas
Lex_4	Abbreviation frequency	ratio of abbreviations to all words
Lex_5	Ratio of long to short words	long words consist of three and more syllables; we identify syllables by vowels and diphthongs
Lex_6	Misspelled words frequency	ratio of misspelled words to all words

Table 1. Features obtained by lexical analysis

ID	Name	Description
Mor_1	Noun frequency	ratio of nouns to all words
Mor_2	Adjective frequency	ratio of adjectives to all words
Mor_3	Pronoun frequency	ratio of pronouns to all words
Mor_4	Verb frequency	ratio of verbs to all words
Mor_5	Proper noun frequency	ratio of proper nouns to all words
Mor_6	Ratio of open to closed word classes	ratio of words that are or are not open to inflection. Open word classes consist of nouns, adjectives, pronouns, numerals and verbs, the rest of word classes are closed
Mor_7	Ratio of functional to content words	ratio of words with meaning to words with only grammatical function; content words include nouns, adjectives, numerals, meaningful (i.e., non-modal) verbs and adverbs, the rest of words are functional
Mor_8	Frequency of sequences of functional words	five or more consecutive functional words with tolerance of one closed word

Table 2. Features obtained by morphological analysis

4.3 Structural Analysis

The main aim of structural analysis is to capture the variability of text distribution and usage of hypertext elements such as links and images. The result of structural analysis is a collection of four features (Table 4).

Besides the first simple structural feature we measure also the standard deviation of section length. Link frequency and image rate are important features especially for distinction between informative and affective articles.

ID	Name	Description
Syn_1	Sentence count	number of identified sentences
Syn_2	Average sentence length	average sentence length in number of words
Syn_3	Ratio of simple to compound sentences	compound sentences consist of two and more sentences
Syn_4	Average subsentence count	subsentence is a simple sentence inside a compound sentence
Syn_5	Ratio of verb to non-verb sentences	sentences with verb to sentences without verb
Syn_6	Dominant tense of predicate candidates	present, future and past
Syn_7	Dominant person of predicate candidates	first, second and third
Syn_8	Dominant number of predicate candidates	singular and plural

Table 3. Features obtained by syntactic analysis

ID	Name	Description
Str_1	Link frequency	ratio of number of links to number of sections
Str_2	Image frequency	ratio of number of images to number of sections
Str_3	Section count	number of sections
Str_4	Standard deviation of section length	deviation of the number of words in sections

Table 4. Features obtained by structural analysis

5 EVALUATION

In order to evaluate our approach, we performed experiments with blog posts in Slovak language. We selected Slovak language as it is our native language and it also represents inflectional languages with strong morphological variability that our method is devised for.

Our aim was to evaluate the proposed features, study performance of different classification algorithms and compare results with state of the art. To evaluate the features, we conducted several experiments focusing on the analysis of correlation between features and classes and evaluating classifier accuracy.

5.1 Dataset

The dataset consists of 578 blogs in Slovak language hosted by the blogging service blog.sme.sk, the biggest blogging service in Slovak republic. The blogging service is not domain specific and both bloggers and readers represent a general audience.

Figure 1. BlogSort web application screenshot. After reading a blog post, a user can decide whether it is informative or affective (bottom-right). The other two columns (for affective dichotomies) appear only after the user selects affective option. To cover the case when the user is uncertain there is always a *hard to say* option in every column. To filter out non-classifiable posts there is a *something different* option in the first column. There was also a motivational factor featured in the application: a score based on number of sorted blog articles by a user was visible to him/her. The presented blog content (left) is in Slovak language.

All the blogs were labelled by several human annotators. In line with our motivation to facilitate navigation of web users in blogosphere (e.g., looking up for blog posts that match their preferences), we employed ordinary web users as annotators. Our aim was to provide independent writing style labels from the proposed classification space and assembly a sufficiently large dataset. To meet this goal, we created BlogSort – a web application allowing participants to read the blog post and decide, whether it is informative or affective. In case of affective blog post they were able to classify it into two other dichotomies. We put stress on gathering the most accurate labels. We did not force the participants to classify the blog if they did not want to do so or if they did not know to select the proper class. The participants had to choose the dominant class, otherwise they could choose *hard to say* option, indicating that it is not possible to decide what the dominant style is. Our aim was to reduce random labels provided by participants at the most and minimize the unwanted noise in the dataset. Figure 1 presents a screenshot of the BlogSort application showing the main user interface.

Together 46 independent participants were involved in blog posts labelling. Figure 2 presents the distribution of readers' classifications among different dichotomies. The obtained results confirm the dominance of affective blogs. Moreover, many of the articles from other categories are photoblogs (often from holidays) and travel guides, which represent a combination of subjective and informative style. Distri-

bution of affective articles into the two proposed dichotomies proved itself to be meaningful and relatively uniform.

All the proposed styles were covered in the collected dataset. For instance, the blog post titled "How Vrbovske vetry blew" represents narrative rational blog about recent attendance (and experience) of the music festival called Vrbovske vetry (can be translated as Winds of Vrbove; Vrbove – the city in Slovakia). The blog post titled "They had gas chambers thought-out really well" about a visit of concentration camp Auschwitz I is an example of narrative emotional blog post. An example of reflective rational post is a post named "Poor situation in the education sector not only in our city" discussing the reasons of low quality of Slovak educational system. Finally, an example representing reflective emotional post is "Smokers. Ignorance? It is time to change that!", where the author wrathfully points to a persisting problem of smoking in public areas.

To determine annotator agreement, we computed Fleiss' kappa for each dichotomy. We employed the variant suitable for multiple ratings per subject with different raters. We obtained kappa equal to 0.446, 0.467, 0.266 for informative/affective, reflective/narrative and emotional/rational dichotomies, respectively. According to the interpretation guidelines, informative/affective and reflective/narrative dichotomies represent moderate agreement (fourth degree on the six degree scale from poor to perfect agreement), while emotional/rational dichotomy represents fair agreement (third degree). Surprisingly, more specific reflective/narrative dichotomy yielded the highest agreement between annotators.

Such agreement suggests that style identification is not an easy task for a human. Blog style may be often not entirely clear mainly if blog posts are longer and tend to include more styles. The authors may choose to mix styles deliberately or it may happen with no intention (the latter applies for reflective or emotional articles). Sometimes the article may contain sum up or conclusion in the opposite style as the rest of the article. Considering the emotional/rational dichotomy, examples of cases difficult for annotation were: the article about female genital mutilation (primarily objective article providing rational arguments against this dreadful ritual, which also contains strong emotional parts), or the political article reflecting on the Velvet revolution in Czechoslovakia (rather emotional reflection on the value of freedom, supported by listing some historical facts).

The acquired set of articles represents a sufficient basis for evaluation of the proposed classification method as well as the correlation between features and classes. The dataset is publicly available for other research groups[1].

Prior to classification of blog posts we undersampled the dataset to adjust the ratio of classes. For each dichotomy, we have chosen approximately the same number of articles from both classes, so that baseline of accuracy in our experiments will always be approximately 50 % (the accuracy of random selection). The Table 5 captures the cardinality of sets of articles for each distribution.

[1] `http://fiit.stuba.sk/~simko/blogsort/dataset.html`

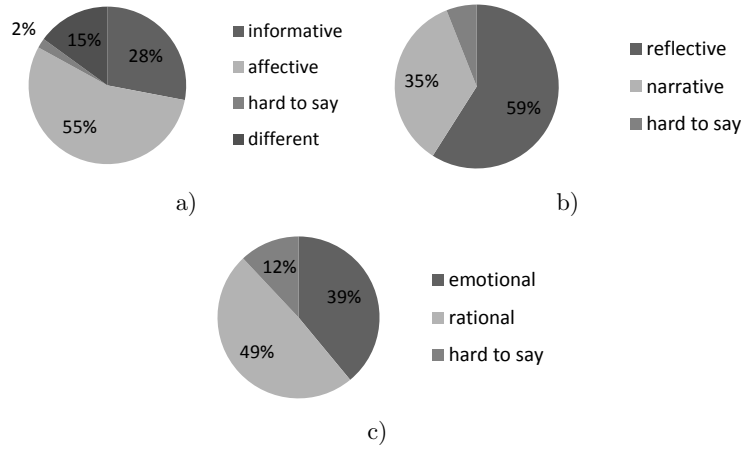

Figure 2. Distribution of readers' classifications a) informative vs. affective, b) reflective vs. narrative, c) emotional vs. rational

	Dichotomy					
	1.		2.		3.	
	informative	affective	reflective	narrative	emotional	rational
Count per class	163	187	183	167	191	229
Total count	350		350		420	

Table 5. Article distribution for each dichotomy

5.2 Feature Extraction

First, we extracted features for each blog article in the obtained dataset by using our implemented feature extractor. It consists of the following parts:

- basic text processing module – basic character- to word-level operations (counting characters, words, computing ratios, etc.).
- morphological annotator – morphological analysis and lemmatization. We utilise morphological and lemma dictionary from the Slovak National Corpus project [29]. For misspelled word detection, we utilize Aspell-sk tool (free Slovak spell-checker)[2].
- lightweight syntactic and structural analyser – our own implementation of basic syntactic and structural analysis focused on sentence-level analysis (sentence types, sentence structure). Considering syntactical analysis, it identifies predicates as basic syntactical units. Considering structural analysis, it performs article-level operations (counting links, images, sections).

[2] `http://www.sk-spell.sk.cx/aspell-sk`

Note that all the proposed features are extracted automatically by our tool.

To evaluate the quantitative performance of our morphological annotator, which is important for correct feature extraction, we performed an experiment involving the created dataset (in total size of about 46 000 words). Our task was to determine the proportion of words that could or could not be annotated by the annotator, i.e., words that can be assigned a morphological tag.

The result of the experiment showed that the annotator was able to recognize 75 % of words, assigning to each at least one lemma and tag. For about 12 % of the rest it could find at least some information, but it could not find their lemma. 13 % of words were classified with wrong spelling. This however does not reveal whether it is a typographical error or a real word, which is just not included in our dictionary. In each article there were approximately 9 words which we knew nothing about. Although the absence of morphological tag may disturb the identification of sentences (especially in case of a verb), the resulting effect on the proposed features is not significant. It may, however, affect existing approaches that are based on counting frequencies of individual words or their lemmas.

5.3 Feature Evaluation

We studied the correlation between features and classes for all three dichotomies. To select the best group of features we used the Correlation-based Feature Selection (CFS) algorithm [9] that aims on selecting group of features with high correlation to class and low correlation between each other.

For the first distribution (informative vs. affective style), a group of nine features was identified by CFS, of which the highest score was obtained by Pronoun frequency (Mor_3) attribute. This finding corresponds with generic approaches to text classification, which deal with similar differentiation (e.g., Sheikha reports informal pronouns to be among the most important features to differentiate between formal and informal texts [30]).

The other features selected by CFS algorithm were Link and Image frequency (Str_1, Str_2), Ratio of long to short words (Lex_5), Ratio of functional to content words (Mor_7), Frequency of sequences of functional words (Mor_8), and all three attributes to measure the properties of a dominant predicate candidate (Syn_6, Syn_7, Syn_8). We can see that syntactical and structural features are important even for basic informative/affective style differentiation that confirms our assumptions.

The distribution of values of top three attributes is depicted in the charts in Figure 3. Dark columns represent informative writing style and light columns affective writing style.

The results lead to the following observations about selected features:

- In affective articles there are more pronouns, short words, functional words and their sequences.
- On the contrary, there is more nouns and shortcuts in informative posts.

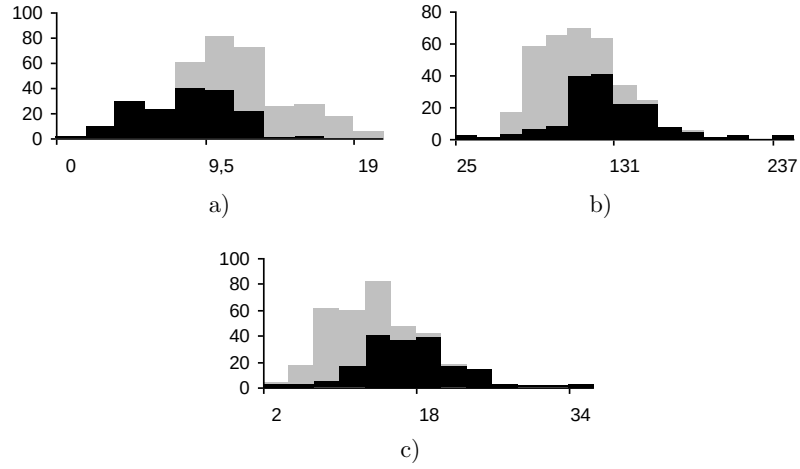

Figure 3. Distribution of attribute values (dark columns: informative style, light columns: affective style): a) Pronoun frequency (Mor_3), b) Ratio of functional to content words (Mor_7), and c) Ratio of long to short words (Lex_5) attributes for first dichotomy.

- The dominance of first singular person of predicate candidate in affective articles confirms the orientation of their text on author's subjective views.
- The ratio of functional to content words acquired a higher score than the ratio of open to closed word classes. Our decision to introduce this new division of words (or word classes) was proved correct.

The group of attributes selected by CFS for second differentiation (reflective vs. narrative style) consisted of five attributes: Dominant person and tense of predicate candidate (Syn_7, Syn_8), Ratio of long to short words (Mor_5), Shortcut frequency (Lex_4) and Standard deviation of section length (Str_4). The obtained results show that:

- In narrative style the first person of predicate candidate prevails, as we expected, since the theme of a narrative text tends to be author's own experience.
- Past tense of predicate candidate, whose dominance we had expected in the narrative text, occurs more frequently in reflective style. By contrast, in the narrative text present and future tense prevail.

We witness that syntactical and structural features are very important and they complement features based on lexical and morphological analysis.

For the third distribution (emotional vs. rational style) a group of 13 attributes was selected by CFS. The best score was obtained by Unique lemmas count (Lex_3) and Word count (Lex_2) attributes. The results show that:

- Articles written in a rational style are mainly consisting of more text (more words, unique lemmas, sentences and words in sentences). They also contain more nouns, verbs and adjectives.
- Surprising is the low score of Standard deviation of section length (Str_4) attribute, which we selected particularly for this dichotomy. It seems that the robustness of the text, which this attribute should capture, is not unique to either one of the categories.

Results show that syntactical and structural features do not have a significant impact in case of recognizing emotional from rational style of blog.

Value distribution of the proposed attributes for each category in all three dichotomies seems to be sufficient for the purposes of classification. An important finding is that there are attributes for each category, which largely distinguish style from one another.

Features proposed in this work are divided into four groups according to the linguistic analysis level they represent: lexical, morphological, syntactic and structural. We have therefore decided to conduct an experiment to verify accuracy of the classifiers for individual groups of features.

We assumed that features based on a higher level of linguistic processing deliver a higher accuracy of classification. We added the fifth group to the four existing ones, which is the result of Correlation-based Feature Selection algorithm, so we were able to evaluate the benefits of feature filtering. The results obtained by using CFS-selected features are compared to the results, wherein all the features were included. Table 6 summarizes the best results obtained for different groups of features.

	informative vs. affective	reflective vs. narrative	emotional vs. rational
Lexical attributes	0.75	0.63	**0.70**
Morphological attributes	0.76	0.63	0.63
Syntactic attributes	0.77	0.71	0.68
Structural attributes	0.67	0.57	0.57
All attributes	**0.83**	**0.73**	0.69
CFS-selected attributes	0.81	0.72	0.69

Table 6. Classification Accuracy. Evaluation of individual feature groups

The obtained results confirmed our assumption only for the first dichotomy, where the accuracy grew with the increasing level of linguistic processing. An exception occurred only in the case of structural attributes. This could be due to the relatively low number of quite general features in this group.

However, our assumption does not apply in two other cases and the increasing level of text processing does not always increase the accuracy of classification. In the third dichotomy we even observe the highest accuracy when applying only lexical features.

Groups of features acquired using CFS algorithm always reached a little lower accuracy than the group containing all features. The proposed feature filtering did not bring an increase in classification accuracy; all attributes seem to be valuable for classification.

5.4 Classifier Evaluation

In order to assess classification accuracy of different classifying algorithms we conducted an experiment where we compared three of them: Naive Bayes (NB; as baseline classifier), Support Vector Machine (SVM) and k-Nearest Neighbours (k-NN). We took advantage of the existing implementation of classification algorithms in the Weka tool [10]. For Support Vector Machine algorithm we selected an implementation based on LibSVM and for k-NN algorithm the IBk implementation. The aim of classifier evaluation is to train selected of-the-shelf classifiers and compare their accuracy. For each classifier we employed 10-fold cross-validation.

We applied all the algorithms for each dichotomy. Table 7 presents the obtained accuracy results.

	Accuracy		
	NB	SVM	k-NN
informative vs. affective	0.80	**0.83**	0.82
reflective vs. narrative	0.65	**0.73**	0.71
emotional vs. rational	0.64	**0.70**	0.67

Table 7. Classifier accuracy evaluation

5.5 Comparison with a Baseline Approach

To identify the contribution of our work clearly, we compared results of our method with the existing method closely related to our work on the same dataset. We drew from an experiment conducted by Elgersma and Rijke, wherein five types of features were proposed to train classifier differentiating between personal and non-personal blogs and were considered as a baseline (term frequency, usage of pronoun, inner and outer links and hosting domain name; cf. Related work section) [8]. In our evaluation we did not consider domain name, which they used as the most useful feature, because we believe that ability to classify blogs regardless of the source is fundamental in heterogeneous blogosphere. Usage of this feature was also not applicable in our case since all blogs in our dataset originate from the same blog hosting service.

Our goal was to apply the baseline method on the original text in Slovak and afterwards on text automatically translated into English (the original language the method was devised for). To translate blog posts in the dataset, we used Google Translate web service. We were able to translate 236 blog articles automatically. The results are shown in Table 8.

	NB	SVM	k-NN
Our method	**0.77**	**0.82**	**0.79**
Baseline method – Slovak	0.76	0.77	0.50
Baseline method – English	0.73	0.62	0.58

Table 8. Classification accuracy. Confrontation with the baseline method [8]

The proposed method achieved the highest accuracy of 82 %. When confronted with the baseline work on the same dataset, it outperformed the method using all classification algorithms. An interesting result came from comparison of original and translated blog articles. We expected that English translation of blog articles in our dataset will perform better, since it is the language the features in [8] were identified for. For example, after translation to English the number of pronouns increases – this better fits classification of the approach [8]. We consider translation quality provided by the service to be sufficient with regards to extracted features as we do not necessarily need the "perfect" translation. We believe that the obtained results suggest that Slovak language is more "suitable" even for work of Elgersma and Rijke. It is due to the way how pronouns are used.

The more important result of the experiment is the knowledge that attributes capturing linguistic features of the text (morphology, syntax, structure) can distinguish between affective and informative blogs better than statistical analysis included in the baseline method. Our work shows that for blog writing style recognition is important to perform deeper linguistic analysis and to consider higher-level specifics of the language family.

6 CONCLUSIONS AND FUTURE WORK

As the amount of user-generated content on the Web grows increasingly, it brings multiple challenges related to the ecosystem of the whole Social Web. They range from content level – where approaches to information processing have to deal with specific vocabulary, language idiosyncrasies, novel short communication forms, with emphasis on multimedia content or other specifics of social content [7] – to social level where researchers measure popularity and influence [42], antisocial behaviour phenomena like trolling [5] or collaborative knowledge sharing [34]. (Irrelevant) Information overload problem is becoming more and more important, and this issue covers blogosphere as well. To allow web users to take full advantage of the Social Web, there have to be continually improved ways how to access to information presented such way effectively. Advancements in information processing are necessary to consider specifics introduced by social media different from traditional web content. To be able to process user-generated content, additional knowledge or metadata about content should be available.

Blogs constitute a web genre that attracts masses of users. In this paper we pointed to the need of fine-grained differentiation of blogs with an aim to capture more characteristics in more detail. We followed known and used differentiation

between informative and affective articles and proposed an additional division of predominant affective type of articles, which better addresses the distribution of blog posts in current blogosphere.

We advocate utilization of comprehensive linguistic analysis for blog writing style detection. The result of our work is a method for recognition and classification of writing style on blogs based on the 26 features considering lexical, morphological, syntactic and structural properties of blog content. To verify the accuracy of our approach we performed an experiment over nontrivial big set of blog articles. We studied the impact of individual types of features and the accuracy of individual classification algorithms. The obtained accuracy is very promising, with a potential of further improvement if the accuracy of feature extraction will be higher. This can be accomplished by further improving linguistic analysers and annotators employed in the pre-processing. Language-specific issues should be considered when dealing with different languages to further extend the feature list.

We confronted our results with the closely related existing approach while concerning blogs classification regardless of the source which is fundamental in heterogeneous blogosphere. The accuracy of our method was higher in all settings, confirming our assumption that classification based on the advanced linguistic structure of the text resulted in increased accuracy of writing style classification on blogs.

The proposed method leaves a space for further improvement especially in the morphological and syntactic analysis. Examining the occurrence of syntactic patterns could contribute to capturing the structure of the text at the sentence level. Fine-tuning or adapting the employed features should bring benefits when applying the method to non-inflected languages. We also plan to research feature personalization for a particular author as the boundaries of each writing style may depend on his/her education, work, social background, political orientation, etc.

Acknowledgments

This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-15-0508, the Scientific Grant Agency of the Slovak Republic, grant No. VG 1/0646/15, and the Ministry of Education, Science, Research and Sport of the Slovak Republic within the Research and Development Operational Programme for the project ITMS 26240220084, co-funded by the European Regional Development Fund.

REFERENCES

[1] Aphinyanaphongs, Y.—Fu, L. D.—Li, Z.—Peskin, E. R.—Efstathiadis, E.—Aliferis, C. F.—Statnikov, A.: A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization. Journal of the Association for Information Science and Technology, Vol. 65, 2014, No. 10, pp. 1964–1987.

[2] Ayyasamy, R. K.: Organizing Information in the Blogosphere: The Use of Unsupervised Approach. International Journal of Soft Computing and Engineering (IJSCE), Vol. 3, 2013, No. 5, pp. 194–198. ISSN 2231-2307.

[3] Bielikova, M.—Kompan, M.—Zelenik, D.: Effective Hierarchical Vector-Based News Representation for Personalized Recommendation. Computer Science and Information Systems, Vol. 9, 2012, No. 1, pp. 303–322.

[4] Blood, R.: We've Got Blog. How blogs Are Changing Our Culture. Perseus Books, New York, 2003.

[5] Cheng, J.—Danescu-Niculescu-Mizil, C.—Leskovec, J.: Antisocial Behavior in Online Discussion Communities. AAAI International Conference on Weblogs and Social Media (ICWSM), 2015.

[6] Dalal, M. K.—Zaveri, M. A.: Automatic Classification of Unstructured Blog Text. Journal of Intelligent Learning Systems and Applications, Vol. 5, 2013, No. 2, pp. 108–114.

[7] Derczynski, L.: Social Media: A Microscope for Public Discourse. Proceedings of the Digital Humanities Congress, 2014.

[8] Elgersma, E.—Rijke, M. D.: Personal vs. Non-Personal Blogs Initial Classification Experiments. Proceedings of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, pp. 723–724.

[9] Hall, A. M.: Correlation-Based Feature Selection for Machine Learning. University of Waikato, Hamilton, New Zealand, 1999.

[10] Hall, A. M.—Frank. E.—Holmes, G.—Pfahringer, B.—Reutemann, P.—Witten, I. H.: The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, Vol. 11, 2009, No. 1, pp. 10–18.

[11] Hoffmann, C. R.: Cohesive Profiling: Meaning and Interaction in Personal Weblogs. John Benjamins Publishing, 2012.

[12] Hoonlor, A.—Szymanski, B. K.—Zaki, M. J.—Chaoji, V.: Document Clustering with Bursty Information. Computing and Informatics, Vol. 31, 2012, No. 6+, pp. 1533–1555.

[13] Husby, S. D.—Barbosa, D.: Topic Classification of Blog Posts Using Distant Supervision. Proceedings of the Workshop on Semantic Analysis in Social Media, ACL, 2012, pp. 28–36.

[14] Karlgren, J.: Textual Stylistic Variation: Choices, Genres and Individuals. In: Argamon, S., Burns, K., Dubnov, S. (Eds.): The Structure of Style. Springer Verlag, 2010, pp. 129–142.

[15] Kompan, M.—Bielikova, M.: News Article Classification Based on a Vector Representation Including Words' Collocations. Advances in Intelligent and Soft Computing. Springer, Vol. 101, 2011, pp. 1–8.

[16] Korenek, P.—Simko, M.: Sentiment Analysis on Microblog Utilizing Appraisal Theory. World Wide Web Journal, Vol. 17, 2014, No. 4, pp. 847–867.

[17] Krishnamurthy, S.: The Multidimensionality of Blog Conversations: The Virtual Enactment of September '11. Association of Internet Researchers Conference 3.0, Maastricht, The Netherlands, 2002.

[18] Kumar, R.—Novak J.—Raghavan P.—Tomkins A.: Structure and Evolution of Blogspace. Communications of the ACM, Vol. 47, 2004, No. 12, pp. 35–39.

[19] Kuzar, T.—Navrat, P.: Preprocessing of Slovak Blog Articles for Clustering. Proceedings of Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, Vol. 3, 2015, pp. 314–317.

[20] Lawson M.: Berners-Lee on the Read/Write Web. BBC, Technology. 2005. Available at: `http://news.bbc.co.uk/1/hi/technology/4132752.stm` (accessed 20/08/2015).

[21] Magdy, W.—Sajjad, H.—El-Ganainy, T.—Sebastiani, F.: Bridging Social Media Via Distant Supervision. arXiv preprint arXiv:1503.04424, 2015.

[22] Mehler, A.—Sharoff, S.—Santini, S.: Genres on the Web: Computational Models and Empirical Studies. Springer Science + Business Media B.V., 2010.

[23] Ni, X.—Xue, G.—Ling, X.—Yu, Y.—Yang, Q.: Exploring in the Weblog Space by Detecting Informative and Affective Articles. Proceedings of 16th International Conference on World Wide Web, 2007, pp. 281–290.

[24] Paltoglou, G.: Sentiment Analysis in Social Media. Online Collective Action. Springer Vienna, 2014, pp. 3–17.

[25] Park, D.: Identifying and Using Formal and Informal Vocabulary. IDP Education, the University of Cambridge and the British Council, the Post Publishing Public Co. Ltd., 2007

[26] Pritsos, D. A.—Stamatatos, E.: Open-Set Classification for Automated Genre Identification. Advances in Information Retrieval. Springer Berlin Heidelberg, 2013, pp. 207–217.

[27] Qu, H.—La Pietra, A.—Poon, S.: Automated Blog Classification: Challenges and Pitfalls. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, 2006, pp. 184–186.

[28] Santini, S.: Characterizing Genres of Web Pages: Genre Hybridism and Individualization. Proceedings of the 40th Hawaii International Conference on System Sciences, 2007, p. 71.

[29] Slovak National Corpus: prim-6.0-public-all. Ľ. Štúr Institute of Linguistics of SAS, Bratislava, 2013. Available at: `http://korpus.juls.savba.sk`.

[30] Sheikha, F. A.—Inkpen, D.: Learning to Classify Documents According to Formal and Informal Style. Linguistic Issues in Language Technology, Vol. 8, 2012, No. 1, pp. 1–29.

[31] Shi, C.—Li, J.—Chen, J.—Chen, X.: Chinese SNS Blog Classification Using Semantic Similarity. Fifth International Conference on Computational Aspects of Social Networks (CASoN), IEEE, 2013, pp. 1–6.

[32] Singh, A. K.—Joshi, R. C.: Semantic Tagging and Classification of Blogs. Proceedings of International Conference on Computer and Communication Technology (ICCCT), IEEE, 2010, pp. 455–459.

[33] Singh, V. K.—Piryani, R.—Uddin, A.—Waila, P.: Sentiment Analysis of Movie Reviews and Blog Posts. 3rd International Advance Computing Conference (IACC 2013), IEEE, 2013, pp. 893–898.

[34] Srba, I.—Bielikova, M.: A Comprehensive Survey and Classification of Approaches for Community Question Answering. ACM Transactions on the Web. Vol. 10, 2016, No. 3, Art. No. 18.

[35] Subramaniyaswamy, V.—Pandian, S. C.: An Improved Approach for Topic Ontology Based Categorization of Blogs Using Support Vector Machine. Journal of Computer Science, Vol. 8, 2012, No. 2, pp. 251–258.

[36] Sugiyanto, S.—Rozi, N. F.—Putri, T. E.—Arifin, A. Z.: Term Weighting Based on Index of Genre for Web Page Genre Classification. JUTI: Scientific Journal of Information Technology, Vol. 12, 2014, No. 1, pp. 27–34.

[37] Vasile, A.—Radulescu, R.—Pavaloiu, I. B.: Topic Classification in Romanian Blogosphere. 12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL), IEEE, 2014, pp. 131–134.

[38] `worldometers.info`: Blogs Written Today. Available at: `http://www.worldometers.info/blogs/` (accessed 20/08/2015).

[39] Xu, R.—Chen, T.—Xia, Y.—Lu, Q.—Liu, B.—Wang, X.: Word Embedding Composition for Data Imbalances in Sentiment and Emotion Classification. Cognitive Computation, Vol. 7, 2015, No. 2, pp. 226–240.

[40] Yatsko, V. A.—Starikov, M. S.—Butakov, A. V.: Automatic Genre Recognition and Adaptive Text Summarization. Automatic Documentation and Mathematical Linguistics, Vol. 44, 2010, No. 3, pp. 111–120.

[41] Zhang, L.—Hu, X.: Word Combination Kernel for Text Classification with Support Vector Machines. Computing and Informatics, Vol. 32, 2013, No. 4, pp. 877–896.

[42] Zhao, W. X.—Liu, J.—He, Y.—Lin, C. Y.—Wen, J. R.: A Computational Approach to Measuring the Correlation Between Expertise and Social Media Influence for Celebrities on Microblogs. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), IEEE, 2014, pp. 460–463.

Martin Virik recieved his Master degree from the Slovak University of Technology in Bratislava in 2011. His research interests include natural language processing, information extraction and classification of user generated content on the web. He currently works as a senior IT specialist at IBM.

Marian Simko received his Ph.D. degree from the Slovak University of Technology in Bratislava in 2012. He is currently Assistant Professor at the same university at the Institute of Informatics, Information Systems and Software Engineering. His research interests include information extraction, ontology engineering, natural language processing, user generated content analysis and technology enhanced learning. He has published more than 30 papers in international journals and conferences related to the topics of his research interests.

Maria Bielikova is Full Professor at the Institute of Informatics, Information Systems and Software Engineering, Slovak University of Technology in Bratislava. Her research interests include web-based systems, emphasizing user modelling and personalization, context awareness, and usability. She received her Ph.D. in computer science from the Slovak University of Technology in Bratislava. She is a member of the Editorial Board of the International Journal of Web Engineering (JWE) and User Modeling and User-Adapted Interaction (UMUAI). She is a senior member of IEEE, a member of the IEEE Computer Society, a senior member of ACM, and a member of the International Society for Web Engineering.