5-15-2019

TOWARDS A TAXONOMY OF TEXT MINING FEATURES

Hansjörg Fromm
Karlsruhe Institute of Technology, hansjoerg.fromm@kit.edu

Thiemo Wambsganss
University of St. Gallen, thiemo.wambsganss@unisg.ch

Matthias Söllner
University of St. Gallen, matthias.soellner@unisg.ch

TOWARDS A TAXONOMY OF TEXT MINING FEATURES

Research in Progress

Hansjörg Fromm, Karlsruhe Institute of Technology, Karlsruhe, Germany,

hansjoerg.fromm@kit.edu

Thiemo Wambsganß, University of St. Gallen, St. Gallen, Switzerland,

thiemo.wambsganss@unisg.ch

Matthias Söllner, University of Kassel, Kassel, Germany, soellner@uni-kassel.de &
University of St. Gallen, St. Gallen, Switzerland, matthias.soellner@unisg.ch

Abstract

Recently, text mining has received special attention from both researchers and practitioners, since it enables the development of intelligent and automated services. Text mining has been influenced by different disciplines like computer science, statistics, computational linguistics and library and information sciences. However, text mining features that evolved in one particular discipline are often unknown or rarely used in the other disciplines. No scientific feature framework exits which facilitates costly feature engineering and evaluation. Therefore, we aim to develop a novel text mining feature taxonomy, which helps researchers and practitioners to develop, refine, compare and evaluate their text mining studies. In this research in progress paper, we focus on laying the foundation for our taxonomy development by presenting our first two research cycles. Here, we were aiming for diversity, not completeness. We derived five dimensions and classified different text features accordingly to provide a deeper understanding.

Keywords: Feature Engineering, Text Mining, Taxonomy, Natural Language Processing.

1 Introduction

Recently, text mining has received special attention from both researchers and practitioners, since it enables the development of intelligent and automated services such as recommender systems, web search, spam detection (Wood, 2016), risk management (Heidinger and Gatzert, 2018), disaster response (Bala et al., 2017), cybercrime prevention (Kontostathis et al., 2010), knowledge discovery (Usai et al., 2018), predictive maintenance (Grabot, 2018) or virtual assistants (Zunic et al., 2016). The main challenge of text mining is to preprocess written text and extract valuable features in a feature vector to enable machine learning (ML) algorithms to reach their maximum performance (Rajman and Vesely, 2004). However, most of the mentioned text mining applications solely rely on the basic feature generation technique *bag-of-words* (BOW) (Joachims, 2002; Nassirtoussi et al., 2014), in which information such as the order and co-occurrence of words are not taken into account. This may lead to an overall underperformance, since copious training data is often not available (Pustejovsky and Stubs, 2013; Bird et al., 2009). As a result, different teams of authors have indicated that enhancing text mining algorithms with the appropriate design, implementation and evaluation of natural language processing (NLP) features - commonly referred to as feature engineering – bears high chances of improving text mining outcomes significantly (Bird et al., 2009; Nassirtoussi et al., 2014; Johnson et al. 2015).

However, extensive feature engineering has a number of challenges. First and foremost, feature engineering currently still depends on human craft rather than on machine learning, since it requires deep domain knowledge to identify and operationalize relevant features. Second, text mining, as Miner et al. (2012) and Talib et al. (2016) point out, has been influenced by different disciplines like *computer science, statistics, computational linguistics,* and *library and information sciences*. Accordingly, text mining features that evolved in one particular discipline are often unknown or rarely used in the other disciplines. Third, no scientific feature framework exits which might help researchers and practitioners to (re)design, compare and evaluate new or existing features across different disciplines or areas of application. Therefore, we aim to develop a novel NLP feature taxonomy, which helps researchers and practitioners to develop, refine, compare and evaluate their text mining studies. Present literature focuses rather on the comparison of particular domain specific features and their learning algorithms but lack a holistic NLP feature framework. Hence, our goal is to develop a comprehensive taxonomy of NLP features based on Nickerson et al.'s (2013) methodological approach. In this research in progress paper, we focus on laying the foundation for our taxonomy development, and on the first two research cycles of our taxonomy development process. Here, we strive to illustrate the diversity of text features, not completeness. In next steps, we aim to investigate a comprehensive taxonomy through providing an overview over all disciplines.

Once our research is completed, we seek to contribute to scientific literature by empirically analyzing the manifold use of NLP features in text mining, and to practice by proposing appropriate dimensions and characteristics for features that might be of value in text mining endeavors. The resulting taxonomy should simplify the comparability of NLP features between different studies, domains or applications and facilitate costly feature engineering of practitioners and scientists. We strive for a set of features diverse enough to demonstrate their commonalties and differences. We believe that if we present this diversity to researchers, they will be inspired and encouraged to use one or the other feature in their own research. Thereby, we want to contribute to a better understanding of NLP feature engineering by answering the following research question:

RQ: *What are the theoretically grounded and empirically validated dimensions and characteristics of NLP features used in text mining?*

In the following, we will firstly introduce the reader to the theoretical background of NLP feature engineering (section 2). In section 3, we present our methodological approach for developing a taxonomy following the work by Nickerson et al. (2013). Afterwards, the results of the first two cycles of the taxonomy development process are presented, followed by an outline of the subsequent steps and the expected implications once our research is completed.

2 Theoretical Background

Text mining uses techniques from two major areas: NLP and ML. Hereby, the main challenge is to preprocess written text and extract valuable features in a feature vector to reach their maximum performance (Rajman and Vesely, 2004) as depicted in *Figure 1*. Features are understood as certain text characteristics or distinct attribute of a text that might bear valuable information for the ML algorithm.

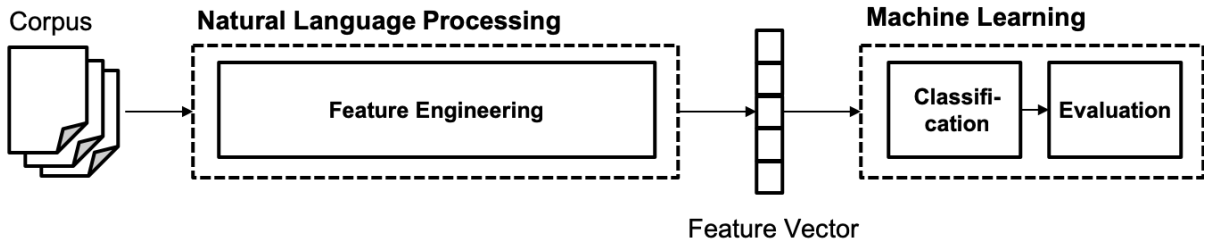

Figure 1. Illustration of feature engineering in the text mining process.

Bird et al., (2009, p. 224) state that "*selecting relevant features and deciding how to encode them for a learning method can have an enormous impact on the learning method's ability to extract a good model*". Thus, most work in building a text classifier is creating relevant features and deciding how to represent them. Bird et al., (2009, p. 224) mention that it is possible to receive decent performance "*by using a fairly simple and obvious set of features …",* however, "*there are usually significant gains to be had by using carefully constructed features based on a thorough understanding of the task at hand.*"

Nevertheless, most features are created through a process of trial-and-error and not by novel feature engineering guided by a taxonomy or framework. In fact, we did not find a taxonomy for text mining features in the literature. What can be found are classifications of text mining features along single dimensions, which are often more a simple categorization of attributes than a set of comprehensive and robust dimensions, e.g., Indurkhya and Damerau, (2010); Johnson et al., (2015) or Missen et al., (2013). Mostly feature engineering is done by intuition about what information might be relevant to the problem. As Bird et al., (2009, p. 224) wrote. "*It's common to start with a "kitchen sink" approach, including all the features that you can think of, and then checking to see which features actually are helpful."*

However, features can be of very different nature, which we want to emphasize with the following examples. The number of occurrences of a specific word within a document is a feature. *Bag-of-words*, as introduced above, records the number of occurrences for every single word in the document. Therefore, BOW is essentially a feature vector or a multi-dimensional feature. Its cardinality is equal to the size of the vocabulary of the document, i.e. the number of different words used. In contrast, there are simple, one-dimensional features like the total number of words, the average sentence length, or the percentage of adjectives appearing in a document.

The different application areas of text analysis and text mining have created a variety of features. *Education and literary sciences* have brought up a number of readability indices which are used to judge the readability level of books and texts. The most widely applied is the Flesch readability index, which is calculated from the number of sentences, words, and syllables of a text (Kincaid, 1975; Flesch, 1943). *Sentiment analysis and opinion mining* (Pang and Lee, 2008) are concerned with the polarity of texts. Polarity is a feature describing the emotion or sentiment present in a text. It can be observed on the word, the sentence, or the document level. Applications *for authorship attribution and verification* analyze the syntactic depth and complexity of texts and have come up with appropriate features like the vocabulary richness or the use of hapax legomena (Stamatatos, 2009; Prasad et al., 2015; Sari et al., 2018). *Requirements analysis* is concerned with the question if textual system requirements are clearly and concisely formulated. Accordingly, features are used that identify weak words, subjunctives, or passive voice, which are all indicators for ambiguous language. *Procedure or instruction mining* tries to find answers on "how to" questions in the vastness of the Internet. This can range from instructions *how to repair an automobile* to recipes *how to prepare a meal*. Typical features used in this domain are the occurrence of enumerations, imperatives, or certain verb-noun combinations.

As depicted, several application areas of text mining exit coming from various disciplines. The disciplines are concerned with different challenges and therefore, develop their own features and their own mode of speaking. A feature framework is required which brings light into the darkness of the various disciplines by illustrating the features' commonalties and differences. However, no comprehensive feature taxonomy exists which depicts the nature of text features across the different disciplines. Hence, we aim to address this literature gap with our stated research question.

3 Method

To answer our research question, we aim to develop a comprehensive taxonomy. Therefore, we follow the method presented by Nickerson et al. (2013), which has been applied by several other studies in the IS field, such as Tan et al. (2016) or Eickhoff et al. (2017). The method follows an iterative and

structured process for developing taxonomies grounded on theoretical foundations (deduction) and empirical evidence (induction) depicted in *Figure 2*. By applying the method of Nickerson et al. (2013), we develop different dimensions and characteristics based on both, published text mining studies and empirical evidence of specific meta attributes. The development of a taxonomy usually starts with defining a specific phenomenon of interest, also called meta-characteristic. The creation of all dimensions and characteristics should be based on contributing to this meta-characteristic. Our meta-characteristic is described by the aim to develop a novel artifact which facilities NLP feature engineering of scientist and practitioners by forming theoretically grounded and empirically validated dimensions and characteristics of NLP features in text mining.

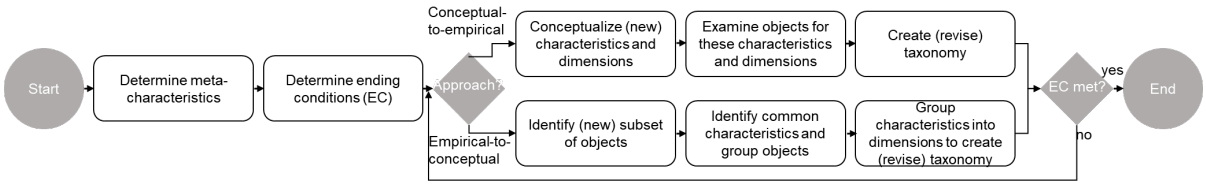

Figure 2. Taxonomy development process based on Nickerson et al. (2013).

Nickerson et al. (2013) suggest different subjective and objective criteria, also called ending conditions, which a taxonomy has to fulfil after the iterative taxonomy development process. We defined the following ending conditions (EC) to determine when to terminate the iterative process.

A) At least one object (text feature) is classified under every characteristic of every dimension.

B) No new dimension or characteristic has been added in the last iteration.

C) Dimensions and characteristics are unique and are not repeated.

D) Every known object (text feature) is classified in the taxonomy.

All ending conditions should be fulfilled by a final taxonomy. However, in the here presented taxonomy not all ending condition are met since we present only intermediate results of our work. Especially condition *D)* is difficult to achieve because it requires an extensive literature study of hundreds of research papers. This will be done in the next phase of our research project. The main goal of the first phase, which we present here, was diversity of features, not completeness. Therefore, we cannot consider the taxonomy development process as completed. However, we believe that we have found a set of features diverse enough to demonstrate their commonalties and differences.
Since we would estimate a high level of knowledge in the research area and multiple text mining studies are available, we conducted a conceptual-to-empirical cycle first, followed by a second empirical-to-conceptual cycle. The development of our taxonomy is illustrated in *Table 1*.

Iteration No.	Approach	Taxonomy	EC met
1	conceptual-to-empirical	T_1 = {Linguistic Category (morphological, lexical, syntactical, semantical), Granularity Level (character, word, sentence, document)}	A, C
2	empirical-to-conceptual	T_2 = {Linguistic Category (morphological, lexical, syntactical, semantical), Granularity Level (character, word, sentence, document), Dimensionality (one-dimensional, multi-dimensional), Representation (binary (presence), integer (count), real number (interval -1, +1), real number (percentage), real number (TF/IDF), real number (general)), Information Source (corpus-based, lexicon-based)}	A, C

Table 1. Taxonomy development iterations.

4 Results

In the following paragraphs, we will state the intermediary results found after conducting two iterations based on Nickerson et al. (2013). First, we will briefly introduce the different dimensions, their characteristics and how we derive them from existing literature. Then, we will illustrate our taxonomy based on feature examples presented in a table to provide a deeper understanding.

A feature can be categorized by the following five dimensions*: dimensionality, representation, linguistic analysis level, granularity level* and *information source*, which are depicted in *Figure 3*.

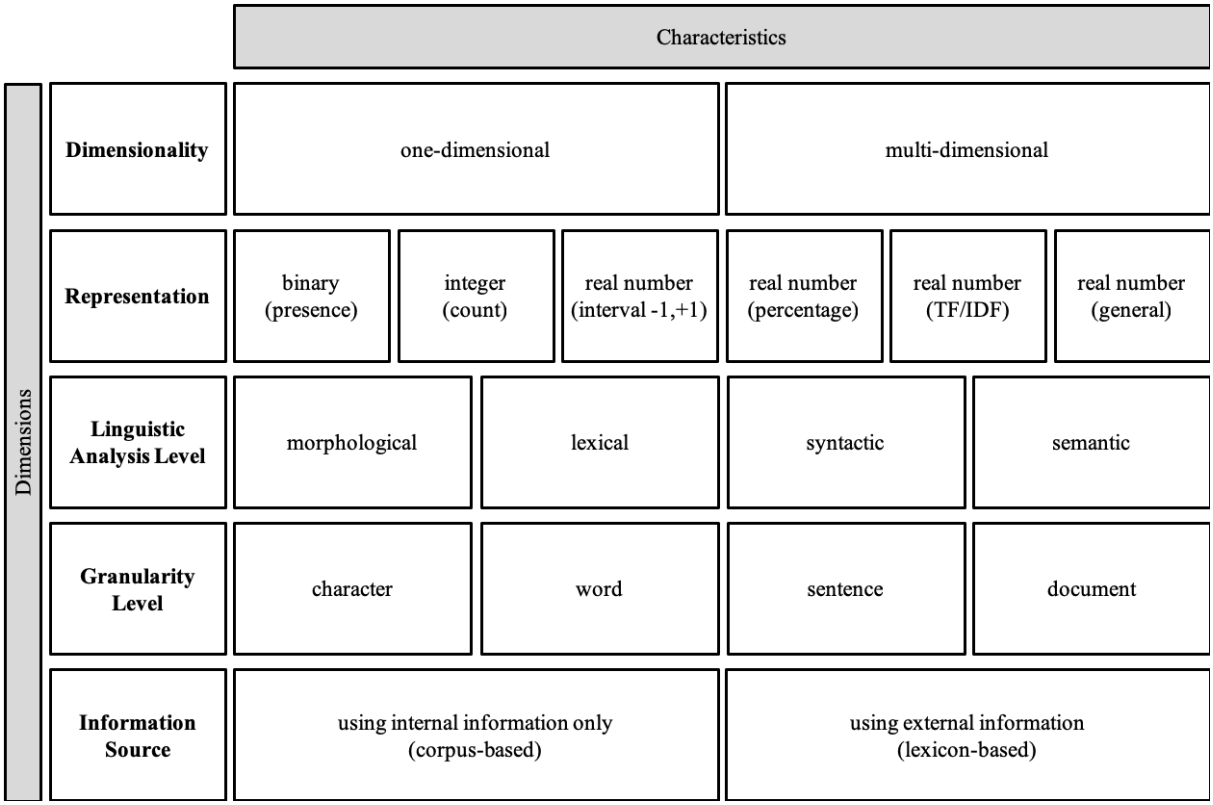

Figure 3. Text feature taxonomy after two iterations based on Nickerson et al. (2013).

Dimensionality of Features

As already stated above, several classifications of text mining features along single dimensions can be found in literature. A dimension that has rarely been brought up in research papers is the dimensionality of features – probably because it is too obvious for most of the authors. There are essentially two cases: features that are expressed in a single number like the number of occurrences of a specific word in a document, the percentage of nouns in a text or the Flesch readability index (e.g., Feng et al., 2010; Kincaid, 1975; Flesch, 1943) and features represented as a vector like *bag-of-words*, *bag-of-n-grams*, *bag-of-POS-grams* (e.g., Palau and Moens, 2009; Brett and Pinna, 2015). The dimensionalities of the latter correspond with the size of the individual vocabulary. Therefore, it is sufficient to distinguish the two characteristics as *one-dimensional* and *multi-dimensional features*.

Representation of Features

All features that we found in the literature were expressed in numbers. They can be distinguished according to their representation (Nassirtoussi et al., 2014) in binary numbers (0, 1), integer numbers (e.g. counts, frequencies), and real numbers (e.g. percentages, values within an interval). *Bag-of-words* appears in three different representations: presence of a word in a document (binary), number of occurrences of a word in a document (*TF* = term frequency), *TFIDF* representation of a word in a document (*TFIDF* = term frequency-inverse document frequency), a real number representing the relative

importance of a word in a document within a given corpus. Since *TFIDF* has a paramount role in text mining, we keep *TFIDF* as a separate class within the dimension "*representation*".

Literature often does not distinguish clearly between the name of a feature and its representation. Some authors call the input of a certain word a feature, some say that the presence or the number of occurrences of a certain work is a feature. Günal et al. (2006) talk about 140 features for spam detection, and they present a list of 140 words. With the representation dimension, things should become clear.

Linguistic Analysis Level

The most popular of the dimensions found in literature are the *"linguistic perspectives"* (Johnson et al., 2015), or in other words, the hierarchy of stages in NLP: *morphological analysis, lexical analysis, syntactical analysis,* and *semantic analysis* (Indurkhya and Damerau, 2010; Johnson et al., 2015). *Lexical analysis* converts a textual input stream into words (also called terms or tokens) that build the vocabulary of the text. *Morphological analysis* looks at the internal structure of the words in more detail (e.g., syllables). *Syntactic analysis* determines the structure of a sentence and the role that each word has within the sentence. *Semantic analysis* is concerned with the meaning of a word, a sentence, or a whole text.

Many authors classify text mining features along these linguistic perspectives. Lexical, syntactic and semantic features are described by Kambhatla (2004) for *relation detection*, by Abbasi et al. (2008) for *sentiment analysis*, by Loni et al. (2011) for *question classification*, by Alzahrani et al. (2012) for *plagiarism detection*. Hancke et al. (2012) use morphological, lexical and semantic features for *readability classification*. Prasad et al. (2015) distinguish lexical, semantic, and structural features for *authorship attribution*. Van der Lee and van den Bosch (2017) describe lexical and syntactic features for *language variety identification*.

The problem that linguistic analysis levels create when developing a feature taxonomy is that there is not necessarily a one-to-one relationship between a feature and an analysis level. *Sentiment analysis* works basically with the meaning of words (with positive or negative sentiments), and thus, requires a lexical and semantic, but not necessarily syntactic analysis (Hatzivassiloglou and Wiebe, 2000). The study of meanings of words is called lexical semantics (Johnson, 2008). The best we can do, to obtain a proper taxonomy dimension for features, is to assign the highest linguistic analysis level to a feature.

Granularity Level

A much less observed characteristic of textual features is their *granularity level* (so called only in Missen et al., 2013; "sub-category" in Suh, 2016). A hierarchy can be described that decomposes a text (a document, a review, a post, a tweet) into sentences which are decomposed in words which are decomposed in characters. Even if the granularity levels seem to be similar to the *linguistic analysis levels*, they are not the same. This becomes vividly illustrated by looking at *unigrams* and *POS-unigrams* (Reyes and Rosso, 2012; Brett and Pinna, 2015). A text *"the highest peak in the country"* (Brett and Pinna, 2015) is composed of the single words (= unigrams) *"the", "highest", "peak", "in", "the", "country*" – which have the syntactic roles (= part-of-speech tags, POS-tags) "*AT0", "AJS", "NN1", "PRP", "AT0", "NN1"*. Occurrences of both *unigrams* and *POS-unigrams* can be counted and thus be used as features. It is clear that *unigrams* require lexical analysis and *POS-unigrams* require syntactic analysis (*"part-of-speech tagging*"). So, *unigrams* and *POS-unigrams* are on different linguistic analysis levels, but they have the same granularity, they are both on word-level. Thus, the granularity level "word" does not necessarily coincide with the linguistic analysis level "lexical". Another example is the semantic feature "*polarity*" (semantic orientation). *Polarity* can be analyzed on the *word level*, on the *sentence level*, or on the *document level* (Missen et al., 2013), which again tells how important the *granularity level* is for differentiation.

		Dimensions				
Feature	**References**	**Dimen-sionality**	**Representation**	**Linguistic Analysis Level**	**Granularity Level**	**Ext. Source**
Frequency of character "@"	Zheng et al. (2006)	one	integer (count)	lexical	character	no
Average number of syllables per word	Feng et al. (2010)	one	real number	morphological	word	no
Average sentence length in words	Suh (2016)	one	real number	lexical	word	no
Bag-of-words (BoW)	many	multi	binary (presence)	lexical	word	no
Bag-of-words (BoW)	many	multi	integer (count)	lexical	word	no
Bag-of-words (BoW)	many	multi	real number (TF/IDF)	lexical	word	no
POS-n-gram vectors	Brett and Pinna (2015); Tang and Cao (2015)	multi	integer (count)	syntactical	word	no
Binary character trigram vectors	Lipka and Stein, (2010)	multi	binary (presence)	lexical	character	no
Readability indices (Flesch, Kincaid, etc.)	Flesch (1943); Kincaid (1975)	one	real number	lexical	document	no
Vocabulary richness (e.g. Yule's K)	Zheng et al. (2006); Suh (2016)	one	real number	lexical	document	no
Fraction of past-tense verbs	Jijkoun et al. (2010)	one	real number (percentage)	syntactical	document	no
Mean number of noun phrases per sentence	Chen and Zechner (2011)	one	real number	syntactical	sentence	no
Mean number of parsing tree levels per sentence	Chen and Zechner (2011) Massung et al. (2013)	one	real number	syntactical	sentence	no
Contextual compatibility score	Liao and Grishman (2010)	one	real number (0; +1)	semantic	word	no
Word polarity	Turney (2002); Rice and Zorn (2013); Agarwal and Mittal (2016)	one	real number (-1; +1)	semantic	word	yes
Sentence polarity	Missen et al. (2013)	one	real number (-1; +1)	semantic	sentence	yes
Document polarity (Review polarity)	Mukherjee and Bhattacharyya (2012)	one	real number (-1; +1)	semantic	document	yes
Bag-of-words with only subjec-tive/objective and positive/negative verbs	Chesley et al. (2006)	multi	integer (count)	semantic	word	yes

Table 2. Feature examples categorized by dimensionality, representation, linguistic analysis level, granularity level and external sources.

Information Source

Semantic features often rely on externally available lexicons that describe the *semantic orientation* or *polarity* (positive or negative) and *subjectivity* (subjective or objective) of individual words or phrases (Taboada et al., 2011). *SentiWordNet* is a such a lexical resource used for *opinion mining* (Baccianella et al., 2010). Since semantic features can also be corpus-based (Liao and Grishman, 2010), a dimension describing the use of external sources is appropriate. Therefore, we included a dimension called "*information source*" with the characteristics *corpus-based* and *lexical-based* to provide further differentiation.

After explaining the dimensions and characteristics of our taxonomy, we want to provide a deeper understanding of the five dimension and their characteristics. Therefore, we allocated different feature examples in the derived dimensions illustrated in *Table 2*.

The five dimensions: *dimensionality*, *representation*, *linguistic analysis level*, *granularity level*, and *use of external sources* were the result of the first two taxonomy development cycles that we conducted according to Nickerson et al. (2013). Not all of the defined ending conditions are satisfied such as that no new dimension or characteristic has been added in the last iteration and that all objects (text features) are categorized. There were a few feature characteristics form text mining studies which do not appear in our taxonomy. Examples are *stylistic features* (Cossu et al., 2015; Mahajan and Zaveri,

2017), *contextual/non-contextual features* (Negi et al., 2014), *frequency-related* and *intensity-related features* (Mahajan and Zaveri, 2017). Careful analysis shows that these characteristics are either application-specific or closely related with characteristics of the already existing dimensions. There is only one dimension that we purposefully did not consider in our taxonomy: the distinction into *textual* and *non-textual features* (Sappelli et al., 2013). *Non-textual features* are features which go beyond pure text analysis. Examples are: the number of links pointing to a web page (Fürnkranz, 1999), the number of clicks on a question and answer pair (Jeon et al., 2006), the number of tweets marked as favorites (Cossu et al., 2015). *Non-textual* features become particularly important if the granularity level considered is above the single document level and looks at collection of documents. This can be hypertext, discussion threads on web forums, or question and answer threads. We decided not to include non-textual features and granularity levels above the document level in our present taxonomy, however, we aim to include these attributes in further iterations.

5 Next Steps & Expected Contributions

In this research in progress paper, we present the initial results of our endeavor towards developing a taxonomy of features used in text mining. As outlined in the beginning of our paper, this research in progress paper focuses on diversity to provide first insights into the depth of the different features. However, in its current form, the taxonomy cannot be considered as complete, meaning that as a next step, we need to dig deeper into the literature in the different domains that apply text mining to make sure that our final taxonomy will capture every relevant feature. This means, we need to engage into a more sophisticated structured approach for identifying the relevant literature (following guidelines, e.g., provided by Webster and Watson (2002) and vom Brocke et al. (2015)) on text mining in the different domains. Based on this enriched database, we need to revise our taxonomy accordingly. Afterwards, we plan to evaluate our taxonomy using semi-structured interviews with text mining researchers and practitioners. Once our research is completed, we contribute to the literature by providing an overview of the different features used in text mining approaches across disciplines so far. Our taxonomy can then serve as a starting point for researchers and practitioners that want to apply text mining, and are looking for potentially relevant features in the feature engineering stage of their project. Additionally, the developed taxonomy will provide researchers a foundation for further theory development and theory testing for feature engineering, and – as intended by Nickerson et al. (2013) – can be extended once new features are developed.

References

Abbasi, A., Chen, H., & Salem, A. (2008). *Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums.* ACM Transactions on Information Systems (TOIS), 26(3), 12.

Agarwal, B., & Mittal, N. (2016). *Prominent feature extraction for sentiment analysis* (pp. 21-45). Cham: Springer.

Alzahrani, S. M., Salim, N., &Abraham, A. (2012). *Understanding plagiarism linguistic patterns, textual features, and detection methods.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(2), 133-149.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.* In Lrec (Vol. 10, No. 2010, pp. 2200-2204).

Bala, M. M., Navya, K., & Shruthilaya, P. (2017). *Text mining on real time Twitter data for disaster response*. Int. J. Civ. Eng. Technol, 8(8), 20-29.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc..

Brett, D., & Pinna, A. (2015). *Patterns, fixedness and variability: using PoS-grams to find phraseologies in the language of travel journalism.* Procedia-Social and Behavioral Sciences, 198, 52-57.

Brink, H., Richards, J. W., & Fetherolf, M. (2017). *Real-world machine learning* (p. 330)

Chen, M., & Zechner, K. (2011, June). *Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech*. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 722-731). Association for Computational Linguistics.

Chesley, P., Vincent, B., Xu, L., & Srihari, R. K. (2006). *Using verbs and adjectives to automatically classify blog sentiment. Training*, 580(263), 233.

Cossu, J. V., Dugué, N., & Labatut, V. (2015, September). *Detecting real-world influence through Twitter.* In: *Network Intelligence Conference (ENIC),* 2015 Second European (pp. 83-90). IEEE.

Eickhoff, M., Muntermann, J. & Weinrich, T. (2017). *What do FinTechs actually do? A Taxonomy of FinTech Business Models.*

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). *A Comparison of Features for Automatic Readability Assessment.* In: pp. 276–284.

Flesch, R. (1943). *Marks of readable style; a study in adult education. Teachers College Contributions to Education.*

Fürnkranz, J. (1999, August*). Exploiting structural information for text classification on the WWW*. In: *International Symposium on Intelligent Data Analysis* (pp. 487-497). Springer, Berlin, Heidelberg.

Geigle, C., Mei, Q., & Zhai, C. (2018). *Feature Engineering for Text Data. Feature Engineering for Machine Learning and Data Analytics*, 15.

Grabot, B. (2018). *Rule mining in maintenance: analysing large knowledge bases*. In: *Computers & Industrial Engineering.*

Günal, S., Ergin, S., Gülmezoğlu, M. B., & Gerek, Ö. N. (2006, September). *On feature extraction for spam e-mail detection*. In: *International Workshop on Multimedia Content Representation, Classification and Security* (pp. 635-642). Springer, Berlin, Heidelberg.

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. *Proceedings of COLING 2012*, 1063-1080.

Hatzivassiloglou, V., & Wiebe, J. M. (2000, July). *Effects of adjective orientation and gradability on sentence subjectivity*. In: *Proceedings of the 18th conference on Computational linguistics*-Volume 1 (pp. 299-305). Association for Computational Linguistics.

Heidinger, D., & Gatzert, N. (2018). *Awareness, determinants and value of reputation risk management: Empirical evidence from the banking and insurance industry*. In: *Journal of Banking & Finance*, 91, 106-118.

Indurkhya, N., & Damerau, F. J. (Eds.). (2010). *Handbook of natural language processing* (Vol. 2). CRC Press.

Jeon, J., Croft, W. B., Lee, J. H., & Park, S. (2006, August). *A framework to predict the quality of answers with non-textual features*. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 228-235). ACM.

Jijkoun, V., de Rijke, M., Weerkamp, W., Ackermans, P., & Geleijnse, G. (2010, June). *Mining user experiences from online forums: an exploration*. In: P*roceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media* (pp. 17-18). Association for Computational Linguistics.

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms* (Vol. 186). Norwell: Kluwer Academic Publishers.

Johnson, K. (2008). *An overview of lexical semantics*. Philosophy Compass, 3(1), 119-134.

Johnson, S. L., Safadi, H., & Faraj, S. (2015). *The emergence of online community leadership. Information Systems Research*, *26*(1), 165-187.

Kambhatla, N. (2004, July). *Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations*. In: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (p. 22). Association for Computational Linguistics.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.*

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). *Text mining and cybercrime*. In: *Text Mining: Applications and Theory*. John Wiley & Sons, Ltd, Chichester, UK.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). *Data mining techniques and applications–a decade review from 2000 to 2011.* In: *Expert Systems with Applications*, vol. 39, no. 12, pp. 11 303–11 311.

Liao, S., & Grishman, R. (2010). *Large Corpus-based Semantic Feature Extraction for Pronoun Coreference.* In: *Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era* (NLPIX 2010) (pp. 60-68).

Lipka, N., & Stein, B. (2010, April). *Identifying featured articles in wikipedia: writing style matters.* In: *Proceedings of the 19th international conference on World wide web* (pp. 1147-1148). ACM.

Loni, B., Van Tulder, G., Wiggers, P., Tax, D. M., & Loog, M. (2011, September). *Question classification by weighted combination of lexical, syntactic and semantic features*. In: *International Conference on Text*, Speech and Dialogue (pp. 243-250). Springer, Berlin, Heidelberg.

Mahajan, R., & Zaveri, M. (2017). *SVNIT $@ $ SemEval 2017 Task-6: Learning a Sense of Humor Using Supervised Approach*. In: *Proceedings of the 11th International Workshop on Semantic Evaluation* (SemEval-2017) (pp. 411-415).

Massung, S., Zhai, C., & Hockenmaier, J. (2013, September). *Structural parse tree features for text representation.* In: *2013 IEEE Seventh International Conference on Semantic Computing* (pp. 9-16). IEEE.

Miner, G., Elder IV, J., & Hill, T. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Missen, M. M. S., Boughanem, M., & Cabanac, G. (2013). *Opinion mining: reviewed from word to document level.* In: *Social Network Analysis and Mining*, *3*(1), 107-125.

Mukherjee, S., & Bhattacharyya, P. (2012). *Sentiment analysis in twitter with lightweight discourse analysis.* In: *Proceedings of COLING 2012*, 1847-1864.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). *Text mining for market prediction: A systematic review*. In*: Expert Systems with Applications*, 41(16), 7653-7670.

Negi, S., & Buitelaar, P. (2014). *INSIGHT galway: syntactic and lexical features for aspect based sentiment analysis*. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 346-350).

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). *A Method for Taxonomy Development and Its Application in Information Systems*. In: *European Journal of Information Systems* (22:3), pp. 336-359.

Palau, R. ; and Moens, M.-F. (2009). *Argumentation Mining: The Detection, Classification and Structure of Arguments in Text*. In: *Proceedings of the 12th international conference on artificial intelligence and law*, pp. 98–107.

Pang, B. and Lee, L. (2008). *Opinion mining and sentiment analysis*. In: *Foundations and Trends in Information Retrieval* 2.12, pp. 1–135.

Prasad, S. N., Narsimha, V. B., Reddy, P. V., & Babu, A. V. (2015). *Influence of lexical, syntactic and structural features and their combination on authorship attribution for Telugu text*. In: *Procedia Computer Science*, 48, 58-64.

Pustejovsky, J. and Stubbs, A. (2013). *Natural Language Annotation*. O'Reilly.

Rajman, M. and Vesely, M. (2004). *From Text to Knowledge: Document Processing and Visualization: a Text Mining Approac*h. In: Springer, Berlin, Heidelberg, pp. 7–24.

Reyes, A., & Rosso, P. (2012). *Making objective decisions from subjective data: Detecting irony in customer reviews*. In: *Decision Support Systems*, 53(4), 754-760.

Rice, D. R., & Zorn, C. (2013). *Corpus-based dictionaries for sentiment analysis of specialized vocabularies*. In: *Proceedings of NDATAD*, 98-115.

Sappelli, M., Verberne, S., & Kraaij, W. (2013). *Combining textual and non-textual features for e-mail importance estimation*.

Sari, Y., Stevenson, M., & Vlachos, A. (2018). *Topic or Style? Exploring the Most Useful Features for Authorship Attribution*. In: *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 343-353).

Scott, S., & Matwin, S. (1999, June). *Feature engineering for text classification*. In: *ICML* (Vol. 99, pp. 379-388).

Stamatatos, E. (2009). *A survey of modern authorship attribution methods*. In: *Journal of the American Society for information Science and Technology*, 60(3), 538-556.

Suh, J. H. (2016). *Comparing writing style feature-based classification methods for estimating user reputations in social media*. Springer, Berlin 5(1), 261.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). *Lexicon-based methods for sentiment analysis*. In: *Computational linguistics*, 37(2), 267-307.

Talib, R., Hanif, Muhammad, K., Ayesha, S., and Fatima, F. (2016). "*Text Mining: Techniques, Applications and Issues"*. In: *Inter- national Journal of Advanced Computer Science and Applications* 7.11, pp. 414– 418.

Tan, C. W., Benbasat, I., and Cenfetelli, R. T. (2016). *An Exploratory Study of the Formation and Impact of Electronic Service Failures*. In: *MIS Quarterly* (40:1), pp. 1-29.

Tang, X., & Cao, J. (2015). *Automatic genre classification via n-grams of part-of-speech tags*. In: *Procedia-Social and Behavioral Sciences*, 198, 474-478.

Turney, P. D. (2002, July). *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*. In: *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417-424). Association for Computational Linguistics.

Usai, A., Pironti, M., Mital, M., & Aouina Mejri, C. (2018). *Knowledge discovery out of text data: a systematic review via text mining*. In: *Journal of Knowledge Management*.

van der Lee, C., & van den Bosch, A. (2017). *Exploring Lexical and Syntactic Features for Language Variety Identification*. In: *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects* (VarDial) (pp. 190-199).

Vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). *Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research.* In: *Communications of the Association for Information Systems*, 37, 205–224.

Webster, J., & Watson, R. T. (2002). *Analyzing the Past to Prepare for the Future: Writing a Literature Review*. In: *MIS Quarterly*, 26(2)

Wood, L. (2016). *Artificial Intelligence (Ai) Market By Technology (Machine Learning, Natural Language Processing (NLP*), Image Processing, And Speech Recognition), Application &amp. In: *Geography - Global Forecast To 2020.* Tech. rep., p. 151.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). *A framework for authorship identification of online messages: Writing-style features and classification techniques*. In*: Journal of the American society for information science and technology*, 57(3), 378-393.

Zunic, E., Djedović, A., & Donko, D. (2016). *Application of Big Data and text mining methods and technologies in modern business analyzing social networks data about traffic tracking*.