What makes a tweet relevant for a topic?

Ke Tao, Fabian Abel, Claudia Hauff, Geert-Jan Houben
Web Information Systems, TU Delft
PO Box 5031, 2600 GA Delft, the Netherlands
{k.tao, f.abel, c.hauff, g.j.p.m.houben}@tudelft.nl

ABSTRACT

Users who rely on microblogging search (MS) engines to find relevant microposts for their queries usually follow their interests and rationale when deciding whether a retrieved post is of interest to them or not. While today's MS engines commonly rely on keyword-based retrieval strategies, we investigate if there exist additional micropost characteristics that are more predictive of a post's relevance and interestingness than its keyword-based similarity with the query. In this paper, we experiment with a corpus of Twitter messages and investigate sixteen features along two dimensions: topic-dependent and topic-independent features. Our in-depth analysis compares the importance of the different types of features and reveals that semantic features and therefore an understanding of the semantic meaning of the tweets plays a major role in determining the relevance of a tweet with respect to a query. We evaluate our findings in a relevance classification experiment and show that by combining different features, we can achieve a precision and recall of more than 35% and 45% respectively.

1. INTRODUCTION

Microblogging services such as Twitter[1] or Sina Weibo[2] have become a valuable source of information particularly for exploring, monitoring and discussing news-related information [7]. Searching for relevant information in such services is challenging as the number of posts published per day can exceed several hundred millions[3].

Moreover, users who search for microposts about a certain topic typically perform a keyword search. Teevan et al. [11] found that keyword queries on Twitter are significantly shorter than those issued for Web search: on Twitter people typically use 1.64 words (or 12.0 characters) to search while on the Web they use, on average, 3.08 words (or 18.8 characters). This can be explained by the length of Twitter messages which is limited to 140 characters so that long queries easily become too restrictive. Short queries on the other hand may result in a large (or too large) number of matching microposts.

For these reasons, building search algorithms that are capable of identifying interesting and relevant microposts for a given topic is a non-trivial and crucial research challenge. In order to take a first step towards solving this challenge, in this paper, we present an analysis of the following question: is a keyword-based retrieval strategy sufficient or can we identify features that are more predictive of a tweet's relevance and interestingness? To investigate this question, we took advantage of last year's TREC[4] 2011 Microblog Track[5], where for the first time an openly accessible search & retrieval Twitter data set with about 16 million tweets was published.

In the context of TREC, the ad-hoc search task on Twitter is defined as follows: given a topic (identified by a title) and a point in time *pt*, retrieve all *interesting and relevant* microposts from the corpus that were posted no later than *pt*. A subset of the tweets that were retrieved by the research groups participating in the benchmark were then judged by human assessors as either relevant to the topic or as non-relevant. For example, "Obama birth certificate" is one of the topics that is part of the TREC corpus. Given the temporal context, one can infer that this topic title refers the discussions about Barack Obama's birth certificate: people were questioning whether Barack Obama was truly born in the United States.

We rely on the judged tweets for our analysis and investigate topic-dependent as well as topic-independent features. Examples of topic-dependent features are the retrieval score derived from retrieval strategies that are based on document and corpus statistics as well as the semantic overlap score which determines the extent of overlap between the semantic meaning of a search topic and a tweet. In addition to these topic-dependent features, we also studied a number of topic-independent features: syntactical features (such as the presence of URLs or hashtags in a tweet), semantic features (such as the diversity of the semantic concepts mentioned in a tweet) and social context features (such as the authority of the user who published the tweet).

The main contributions of our work can be summarized as follows:

- We present a set of strategies for the extraction of fea-

[1] http://twitter.com/
[2] http://www.weibo.com/
[3] http://blog.twitter.com/2011/06/200-million-tweets-per-day.html

[4] http://trec.nist.gov/
[5] http://sites.google.com/site/trecmicroblogtrack/

Published as part of the #MSM2012 Workshop proceedings, available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

tures from Twitter messages that allow us to predict the relevance of a post for a given topic.
- Given a set of more than 38,000 tweets that were manually labeled as relevant or not relevant for a set of 49 topics, we analyze the features and characteristics of relevant and interesting tweets.
- We evaluate the effectiveness of the different features for predicting the relevance of tweets for a topic and investigate the impact of the different features on the quality of the relevance classification. We also study to what extent the success of the classification depends on the type of topics (e.g. topics of short-term vs. topics of long-term interest) for which relevant tweets should be identified.

2. RELATED WORK

Since its launch in 2006 Twitter attracted a lot of attention, both in the general public as well as in the research community. Researchers started studying microblogging phenomena to find out what kind of information is discussed on Twitter [7], how trends evolve on Twitter [8], or how one detects influential users on Twitter [12]. Applications have been researched that utilize microblogging data to enrich traditional news media with information from Twitter [6], to detect and manage emergency situations such as earthquakes [10] or to enhance search and ranking of Web sites which possibly have not been indexed yet by Web search engines.

So far, search on Twitter or other microblogging platforms such as Sina Weibo has not been studied extensively. Teevan et al. [11] compared the search behavior on Twitter with traditional Web search behavior. It was found that keyword queries that people issue to retrieve information from Twitter are, on average, significantly shorter than queries submitted to traditional Web search engines (1.64 words vs. 3.08 words). This finding indicates that there is a demand to investigate new algorithms and strategies for retrieving relevant information from microblogging streams.

Bernstein et al. [2] proposed an interface that allows for exploring tweets by means of tag clouds. However, their interface is targeted towards browsing the tweets that have been published by the people whom a user is following and not for searching the entire Twitter corpus. Jadhav et al. [6] developed an engine that enriches the semantics of Twitter messages and allows for issuing SPARQL queries on Twitter streams. In previous work, we followed such a semantic enrichment strategy to provide faceted search capabilities on Twitter [1]. Duan et al. [5] investigated features such as Okapi BM25 relevance scores or Twitter specific features (length of a tweet, presence or absence of a URL or hashtag, etc.) in combination with RankSVM to learn a ranking model for tweets (learning to rank). In an empirical study, they found that the length of a tweet and information about the presence of a URL in a tweet are important features to rank relevant tweets. In this paper, we re-visit some of the features proposed by Duan et al. [5] and introduce novel semantic measures that allow us to estimate whether a micropost is relevant to a given topic or not.

3. FEATURES OF MICROPOSTS

In this section, we provide an overview of the different features that we analyze to estimate the relevance of a Twitter message to a given topic. We present topic-sensitive features that measure the relevance with respect to the topic (keyword-based and semantic-based relevance) and topic-insensitive measures that do not consider the actual topic but solely exploit syntactical or semantic tweet characteristics. Finally, we also consider contextual features that, for example, characterize the creator of a tweet.

3.1 Keyword-based Relevance Features

keyword-based relevance score (Indri-based query relevance): To calculate the retrieval score for pair of (topic, tweet), we employ the language modeling approach to information retrieval [13]. A language model θ_t is derived for each document (tweet). Given a query Q with terms $Q = \{q_1, ..., q_n\}$ the document language models are ranked with respect to the probability $P(\theta_t|Q)$, which according to the Bayes theorem can be expressed as:

$$P(\theta_t|Q) = \frac{P(Q|\theta_t)P(\theta_t)}{P(Q)} \quad (1)$$

$$\propto P(\theta_t) \prod_{q_i \in Q} P(q_i|\theta_t). \quad (2)$$

This is the standard query likelihood based language modeling setup which assumes term independence. Usually, the prior probability of a tweet $P(\theta_t)$ is considered to be uniform, that is, each tweet in the corpus is equally likely. The language models are multinomial probability distributions over the terms occurring in the tweets. Since a maximum likelihood estimate of $P(q_i|\theta_t)$ would result in a zero probability of any tweet that misses one or more of the query terms in Q, the estimate is usually smoothed with a background language model, generated over all tweets in the corpus. We employed Dirichlet smoothing [13]:

$$P(q_i|\theta_t) = \frac{c(q_i, t) + \mu P(q_i|\theta_C)}{|t| + \mu}. \quad (3)$$

Here, μ is the smoothing parameter, $c(q_i, t)$ is the count of term q_i in t and $|t|$ is the length of the tweet. The probability $P(q_i|\theta_C)$ is the maximum likelihood probability of term q_i occurring in the collection language model θ_C (derived by concatenating all tweets in the corpus).

Due to the very small probabilities of $P(Q|\theta_t)$, we utilize $\log(P(Q|\theta_t))$ as feature scores. Note that this score is always negative. The greater the score (that is, the less negative), the more relevant the tweet is to the query.

3.2 Semantic-based Relevance Features

semantic-based relevance score This feature is also a retrieval score calculated according to Section 3.1 though with a different set of queries. Since the average length of search queries submitted to microblog search engines is lower than in traditional Web search, it is necessary to understand the information need behind the query. The search topics provided as part of the TREC data set contain abbreviations, part of names, and nicknames. One example (cf. Table 1) is the first name "Jintao" (in the query: "Jintao visit US") which refers to the President of the People's Republic of China. However, in tweets he is also referred to as "President Hu", "Chinese President", etc. If these semantic variants of a person's name and titles would be considered when deriving an expanded query, a wider variety of potentially relevant tweets could be found. We utilize the well-known Named-Entity-Recognition (NER) service DBPedia

Query	**Jintao** visits US	
Entity	**Annotated Text**	**Possible Concepts**
Hu Jintao	Jintao	Hu, Jintao, Hu Jintao

Table 1: Example of entity recognition and possible concepts in the query

Spotlight[6] to identify names and their synonyms in the original query. We merge the found concepts into an expanded query which is then used as input to the retrieval approach described earlier.

isSemanticallyRelated It is a boolean value that shows whether there is a semantic overlap between the topic and the tweet. This requires us to employ DBpedia Spotlight on the topic as well as the tweets. If there is an overlap in the identified DBpedia concepts, the value of this feature is *true*, otherwise it is *false*.

3.3 Syntactical Features

Syntactical features describe elements that are mentioned in a Twitter message. We analyze the following properties:

hasHashtag This is a boolean property which indicates whether a given tweet contains at least one hashtag or not. Twitter users typically apply hashtags in order to facilitate the retrieval of the tweet. For example, by using a hashtag people can join a discussion on a topic that is represented via that hashtag. Users, who monitor the hashtag, will retrieve all tweets that contain it. Teevan et al. [11] showed that such monitoring behavior is a common practice on Twitter to retrieve relevant Twitter messages. Therefore, we investigate whether the occurrence of hashtags (possibly without any obvious relevance to the topic) is an indicator for the relevance and interestingness of a tweet.
Hypothesis H1: tweets that contain hashtags are more likely to be relevant than tweets that do not contain hashtags.

hasURL Dong et al. [4] showed that people often exchange URLs via Twitter so that information about trending URLs can be exploited to improve Web search and particularly the ranking of recently discussed URLs. Therefore, the presence of a URL (boolean property) can be an indicator for the relevance of a tweet.
Hypothesis H2: tweets that contain a URL are more likely to be relevant than tweets that do not contain a URL.

isReply On Twitter, users can reply to the tweets of other people. This type of communication can, for example, be used to comment on a certain message, to answer a question or to chat with other people. Chen et al. [3] studied the characteristics of reply chains and discovered that one can distinguish between users who are merely interested in news-related information and users who are also interested in social chatter. For deciding whether a tweet is relevant for a news-related topic, we therefore assume that the boolean *isReply* feature, which indicates whether a tweet is a reply to another tweet, can be a valuable signal.
Hypothesis H3: tweets that are formulated as a reply to another tweet are less likely to be relevant than other tweets.

length The length of a tweet—measured in the number of characters—may also be an indicator for the relevance or interestingness. We hypothesize that the length of a Twitter message correlates with the amount of information that is conveyed in the message.
Hypothesis H4: the longer a tweet, the more likely it is to be relevant and interesting.

The values of boolean properties are set to 0 (false) and 1 (true) while the length of a Twitter message is measured by the number of characters divided by 140 which is the maximum length of a Twitter message.

There are further syntactical features that can be explored such as the mentioning of certain character sequences including emoticons, question marks, exclamation marks, etc. In line with the *isReply* feature, one could also utilize knowledge about the re-tweet history of a tweet, e.g. a boolean property that indicates whether the tweet is a copy from another tweet or a numeric property that counts the number of users who re-tweeted the message. However, in this paper we are merely interested in original messages that have not been re-tweeted yet[7] and therefore also merely in features which do not require any knowledge about the history of a tweet. This allows us to estimate the relevance of a message as soon as it is published.

3.4 Semantic Features

In addition to the semantic relevance scores described in Section 3.2, one can also analyze the semantics of a Twitter message independently from the topic of interest. We therefore utilize again the DBpedia entity extraction provided by DBpedia Spotlight to extract the following features:

#entities The number of DBpedia entities that are mentioned in a Twitter message may give further evidence about the potential relevance and interestingness of a tweet. We assume that the more entities can be extracted from a tweet, the more information it contains and the more valuable it is. For example, in the context of the discussion about birth certificates we find the following two tweets in our dataset:

t_1: "Despite what her birth certificate says, my lady is actually only 27"
t_2: "Hawaii (Democratic) lawmakers want release of Obama's birth certificate"

When reading the two tweets, without having a particular topic or information need in mind, it seems that t_2 has a higher likelihood to be relevant for some topic for the majority of the Twitter users than t_1 as it conveys more entities that are known to the public and available on Wikipedia and DBpedia respectively. In fact, the entity extractor is able to detect one entity, *db:Birth_certificate*, for tweet t_1 while it detects three additional entities for t_2: *db:Hawaii*, *db:Legislator* and *db:Barack_Obama*.
Hypothesis H5: the more entities a tweet mentions, the more likely it is to be relevant and interesting.

#entities(type) Similarly to counting the number of entities that occur in a Twitter message, we also count the number of entities of specific types. The rationale behind this feature being that some types of entities might be a stronger indicator for relevance than others. The importance of a specific entity type may also depend on the topic.

[6]DBpedia Spotlight, `http://spotlight.dbpedia.org/`

[7]This is in line with the relevance judgments provided by TREC which did not consider re-tweeted messages.

For example, when searching for Twitter messages that report about wild fires in a specific area, location-related entities may be more interesting than product-related entities. In this paper, we count the number of entity occurrences in a Twitter message for five different types: locations, persons, organizations, artifacts and species (plants and animals).
Hypothesis H6: different types of entities are of different importance for estimating the relevance of a tweet.

diversity The diversity of semantic concepts mentioned in a Twitter message can also be exploited as an indicator for the potential relevance and interestingness of a tweet. We therefore count the number of distinct types of entities that are mentioned in a Twitter message. For example, for the two tweets t_1 and t_2 mentioned earlier, the diversity score would be 1 and 4 respectively as for t_1 only one type of entity is detected (*yago:PersonalDocuments*) while for t_2 also instances of *db:Person* (person), *db:Place* (location) and *owl:Thing* (the role *db:Legislator* is not further classified) are detected.
Hypothesis H7: the greater the diversity of concepts mentioned in a tweet, the more likely it is to be interesting and relevant.

sentiment Naveed et al. [9] showed that tweets which contain negative emoticons are more likely to be re-tweeted than tweets which feature positive emoticons. The sentiment of a tweet may thus impact the perceived relevance of a tweet. Therefore, we classify the the semantic polarity of a tweet into positive, negative or neutral using *Twitter Sentiment*[8].
Hypothesis H8: the likelihood of a tweet's relevance is influenced by its sentiment polarity.

3.5 Contextual Features

In addition to the aforementioned features, which describe characteristics of the Twitter messages, we also investigate features that describe the context in which a tweet was publish. In our analysis, we investigate the social and temporal context:

social context The social context describes the creator of a Twitter message. Different characteristics of the message creator may increase or decrease the likelihood of her tweets being relevant and interesting such as the number of followers or the number of tweets from this user that have been re-tweeted. In this paper, we apply a light-weight measure to characterize the creator of a message: we count the number of tweets which the user has published.
Hypothesis H9: the higher the number of tweets that have been published by the creator of a tweet, the more likely it is that the tweet is relevant.

temporal context The temporal context describes *when* a tweet was published. The creation time can be specified with respect to the time when a user is requesting tweets about a certain topic (query time) or it can be independent of the query time. For example, one could specify at which hour during the day the tweet was published or whether it was created during the weekend. In our analysis, we utilize the temporal distance (in seconds) between the query time and the creation time of the tweet. *Hypothesis H10: the lower the temporal distance between the query time and the creation time of a tweet, the more likely is the tweet relevant to the topic.*

[8] `http://twittersentiment.appspot.com/`

Contextual features may also refer to characteristics of Web pages that are linked from a Twitter message. For example, one could exploit the PageRank scores of the referenced Web sites to estimate the relevance of a tweet or one could categorize the linked Web pages to discover the types of Web sites that usually attract attention on Twitter. We leave the investigation of such additional contextual features for future work.

4. FEATURE ANALYSIS

In this section, we describe and characterize the Twitter corpus with respect to the features that we presented in the previous section.

4.1 Dataset Characteristics

We use the Twitter corpus which was used in the microblog track of TREC 2011[9]. The original corpus consists of approximately 16 million tweets, posted over a period of 2 weeks (January 24 until February 8th, inclusive). We utilized an existing language detection library[10] to identify English tweets and found that 4,766,901 tweets were classified as English. Employing NER on the English tweets resulted in a total over six million named entities among which we found approximately 0.14 million distinct entities. Besides the tweets, 49 topics were given as the targets of retrieval. TREC assessors judged the relevance of 40,855 topic-tweet pairs which we use as ground truth in our experiments. 2,825 tweets were judged as relevant for a given topic while the majority of the tweet-topic pairs (37,349) were marked as non-relevant.

4.2 Feature Characteristics

In Table 2 we list the average values and the standard deviations of the features and the percentages of true instances for boolean features respectively. It shows that relevant and non-relevant tweets show, on average, different characteristics for several features.

As expected, the average keyword-based relevance score of tweets, which are judged as relevant for a given topic, is much higher than the one for non-relevant tweets: -10.709 in comparison to -14.408 (the higher the value the better, see Section 3.1). Similarly, the semantic-based relevance score, which exploits the semantic concepts mentioned in the tweets (see Section 3.2) while calculating the retrieval rankings, shows the same characteristic. The isSemanticallyReleated feature, which is a binary measure of the overlap between the semantic concepts mentioned in the query and the respective tweets, is also higher for relevant tweets than for non-relevant tweets. Hence, when we consider the topic-dependent features (keyword-based and semantic-based), we find first indicators that the hypotheses behind these features hold.

For the syntactical features we observe that, regardless of whether the tweets are relevant to a topic or not, the ratios of tweets that contain hashtags are almost the same (about 19%). Hence, it seems that the presence of a hashtag is not necessarily an indicator for relevance. However, the presence of a URL is potentially a very good indicator: 81.9% of the relevant tweets feature a URL whereas only 54.1% of the non-relevant tweets contain a URL. A possible explana-

[9] `http://trec.nist.gov/data/tweets/`
[10] Language detection, `http://code.google.com/p/language-detection/`

Category	Feature	Relevant	Standard deviation	Non-relevant	Standard deviation
keyword relevance	keyword-based	-10.709	3.5860	-14.408	2.6442
semantic relevance	semantic-based	-10.308	3.7363	-14.264	3.1872
	isSemanticallyRelated	25.3%	43.5%	4.6%	22.6%
syntactical	hasHashtag	19.1%	39.2%	19.3%	39.9%
	hasURL	81.9%	38.5%	54.1%	49.5%
	isReply	3.4%	18.0%	14.2%	34.5%
	length (in characters)	90.323	30.81	87.797	36.17
semantics	#entities	2.367	1.605	1.880	1.777
	#entities(person)	0.276	0.566	0.188	0.491
	#entities(organization)	0.316	0.589	0.181	0.573
	#entities(location)	0.177	0.484	0.116	0.444
	#entities(artifact)	0.188	0.471	0.245	0.609
	#entities(species)	0.005	0.094	0.012	0.070
	diversity	0.795	0.788	0.597	0.802
	sentiment (-1=neg, 1=pos)	-0.025	0.269	0.042	0.395
contextual	social context (#tweets by creator)	12.287	19.069	12.226	20.027
	temporal context (time distance in days)	4.85	4.48	3.98	5.09

Table 2: The comparison of features between relevant tweets and non-relevant tweets

tion for this difference is that the tweets containing URLs tend to feature also an attractive short title, especially for breaking news, in order to attract people to follow the link. Moreover, the actual content of the linked Web site may also stipulate users when assessing the relevance of a tweet. In Hypothesis 3 (see Section 3.3), we speculate that messages which are replies to other tweets are less likely to be relevant than other tweets. The results listed in Table 2 support this hypothesis: only 3.4% of the relevant tweets are replies in contrast to 14.2% of the non-relevant tweets. The length of the tweets that are judged as relevant is, on average, 90.3 characters, which is slightly longer than for the non-relevant ones (87.8 characters).

The comparison of the topic-independent semantic features also reveals some differences between relevant and non-relevant tweets. Overall, relevant tweets contain more entities (2.4) than non-relevant tweets (1.9). Among the five most frequently mentioned types of entities, persons, organizations, and locations occur more often in relevant tweets than in non-relevant ones. On average, messages are therefore considered as more likely to be relevant or interesting for users if they contain information about people, involved organizations, or places. Artifacts (e.g. tangible things, software) and species (e.g. plants, animals) are more frequent in non-relevant tweets. However, counting the number of entities of type species seems to be a less promising feature since the fraction of tweets which mention a species is fairly low.

The diversity of content mentioned in a Twitter message—i.e. the number of distinct types (only person, organization, location, artifact, and species are considered)—is potentially a good feature: the semantic diversity is higher for the relevant tweets (0.8) than for the non-relevant ones (0.6). In addition to the entities that are mentioned in the tweets, we also conducted a sentiment analysis of the tweets (see Section 3.4). Although most of the tweets are neutral (sentiment score = 0), the average sentiment score for relevant tweets is negative (-0.025). This observation is in line with the finding made by Naveed et al. [9] who found that negative tweets are more likely to be re-tweeted.

Finally, we also attempted to determine the relationship between a tweet's likelihood of relevance and its context. With respect to the social context, we however do not observe a significant difference between relevant an non-relevant tweets: users who publish relevant tweets are, on average, not more active than publishers of non-relevant tweets (12.3 vs. 12.2). For the temporal context, the average distance between the time when a user requests tweets about a topic and the creation time of tweets is 4.85 days for relevant tweets and 3.98 for non-relevant tweets. However, the standard deviations of these scores is with 4.53 days (relevant) and 4.39 days (non-relevant) fairly high. This indicates that the temporal context is not a reliable feature for our dataset. Preliminary experiments indeed confirmed the low utility of the temporal feature. However, this observation seems to be strongly influenced by the TREC dataset itself which was collected within a short time period of time (two weeks). In our evaluations, we therefore do not consider the temporal context and leave an analysis of the temporal features for future work.

5. EVALUATION OF FEATURES FOR RELEVANCE PREDICTION

Having analyzed the dataset and the proposed features, we now evaluate the quality of the features for predicting the relevance of tweets for a given topic. We first outline the experimental setup before we present our results and analyze the influence of the different features on the performance for the different types of topics.

5.1 Experimental Setup

We employ logistic regression to classify tweets as relevant or non-relevant to a given topic. Due to the small size of the topic set (49 topics), we use 5-fold cross validation to evaluate the learned classification models. For the final setup, 16 features were used as predictor variables (all features listed in Table 2 except for the temporal context). To conduct our experiments, we rely on the machine learning toolkit Weka[11]. As the number of relevant tweets is considerably smaller than the number of non-relevant tweets, we employed a cost-sensitive classification setup to prevent the classifier from following a best match strategy where simply all tweets are marked as non-relevant. As the estimation for the negative class achieves a precision and recall both over 90%, we focus on the precision and recall of the relevance classification (the positive class) in our evaluation as we aim to investigate the characteristics that make tweets relevant to a given topic.

[11] http://www.cs.waikato.ac.nz/ml/weka/

Features	Precision	Recall	F-Measure
keyword relevance	0.3040	0.2924	0.2981
semantic relevance	0.3053	0.2931	0.2991
topic-sensitive	0.3017	0.3419	0.3206
topic-insensitive	0.1294	0.0170	0.0300
without semantics	0.3363	0.4828	0.3965
all features	0.3674	0.4736	0.4138

Table 3: Performance results of relevance predictions for different sets of features.

Feature Category	Feature	Coefficient
keyword-based	keyword-based	0.1701
semantic-based	semantic-based	0.1046
	isSemanticallyRelated	0.9177
syntactical	hasHashtag	0.0946
	hasURL	1.2431
	isReply	-0.5662
	length	0.0004
semantics	#entities	0.0339
	#entities(person)	-0.0725
	#entities(organization)	-0.0890
	#entities(location)	-0.0927
	#entities(artifact)	-0.3404
	#entities(species)	-0.5914
	diversity	0.2006
	sentiment	-0.5220
contextual	social context	-0.0042

Table 4: The feature coefficients were determined across all topics. The total number of topics is 49. The three features with the highest absolute coefficient are underlined.

5.2 Influence of Features on Relevance Prediction

Table 3 shows the performances of estimating the relevance of tweets based on different sets of features. Learning the classification model solely based on the keyword-based or semantic-based relevance scoring features leads to an F-Measure of 0.2981 and 0.2991 respectively. There is thus no notable difference between the two topic-sensitive features. However, by combining both features (see topic-sensitive in Table 3) the F-Measure increases which is caused by a higher recall, increasing from 0.29 to 0.34. It appears that the keyword-based and semantic-based relevance scores complement each other.

As expected, when solely learning the classification model based on the topic-independent features—i.e. without measuring the relevance to the given topic—the quality of the relevance prediction is poor. The best performance is achieved when all features are combined. A precision of 36.74% means that more than a third of all tweets that our approach classifies as relevant are indeed relevant, while the recall level (47.36%) implies that our approach discovers nearly half of all relevant tweets. Since microblog messages are very short, a significant number of tweets can be read quickly by a user when presented in response to her search request. In such a setting, we believe such a classification accuracy to be sufficient. Overall, the semantic features seem to play an important role as they lead to a performance improvement with respect to the F-Measure from 0.3965 to 0.4138. We will now analyze the impact of the different features in detail.

One of the advantages of the logistic regression model is, that it is easy to determine the most important features of the model by considering the absolute weights assigned to them. For this reason, we have listed the relevant-tweet prediction model coefficients for all employed features in Table 4. The features influencing the model the most are:

- *hasURL*: Since the feature coefficient is positive, the presence of a URL in a tweet is more indicative of relevance than non-relevance. That means, that hypothesis H2 (Section 3.3) holds.
- *isSemanticallyRelated*: The overlap between the identified DBpedia concepts in the topics and the identified DBpedia concepts in the tweets is the second most important feature in this model. This is an interesting observation, especially in comparison to the keyword-based relevance score, which is only the ninth important feature among the evaluated ones. It implies that a standard keyword-based retrieval approach, which performs well for longer documents, is less suitable for microposts.
- *isReply*: This feature, which is *true* (= 1) if a tweet is written in reply to a previously published tweet has a negative coefficient which means that tweets which are replies are less likely to be in the relevant class than tweets which are not replies, confirming hypothesis H3 (Section 3.3).
- *sentiment*: The coefficient of the sentiment feature is similarly negative, which suggests that a negative sentiment is more predictive of relevance than a positive sentiment, in line with our hypothesis H8 (Section 3.4).

We note that the keyword-based similarity, while being positively aligned with relevance, does not belong to the most important features in this model. It is superseded by syntactic as well as semantic-based features. When we consider the non-topical features only, we observe that interestingness (independent of a topic) is related to the potential amount of additional information (i.e. the presence of a URL), the clarity of the tweet overall (a tweet in reply may be only understandable in the context of the contextual tweets) and the different aspects covered in the tweet (as evident in the diversity feature). It should also be pointed out that the negative coefficients assigned to most topic-insensitive entity count features (*#entities(X)*) is in line with the results in Table 2.

5.3 Influence of Topic Characteristics on Relevance Prediction

In all reported experiments so far, we have considered the entire set of topics available to us. In this section, we investigate to what extent certain topic characteristics play a role for relevance prediction and to what extent those differences lead to a change in the logistic regression models.

Consider the following two topics: *Taco Bell filling lawsuit* (MB020[12]) and *Egyptian protesters attack museum* (MB010). While the former has a business theme and is likely to be mostly of interest to American users, the latter topic belongs into the politics category and can be considered as being of global interest, as the entire world was watching the events in Egypt unfold. Due to these differences we defined a number of topic splits. A manual annotator then decided for each split dimension into which category the topic should fall. We investigated four topic splits, three splits with two

[12]The identifiers of the topics correspond to the ones used in the official TREC dataset.

Performance	Measure	popular	unpopular	global	local	persistent	occasional
	#topics	24	25	18	31	28	21
	#samples	19803	21052	16209	25646	22604	18251
	precision	0.3596	0.3579	0.3442	0.3726	0.3439	0.4072
	recall	0.4308	0.5344	0.4510	0.4884	0.4311	0.5330
	F-measure	0.3920	0.4287	0.3904	0.4227	0.3826	0.4617
Feature Category	**Feature**	**popular**	**unpopular**	**global**	**local**	**persistent**	**occasional**
keyword-based	keyword-based	0.1018	0.2475	0.1873	0.1624	0.1531	0.1958
semantic-based	semantic-based	0.1061	0.1312	0.1026	0.1028	0.0820	0.1560
	isSemanticallyRelated	1.1026	0.5546	0.9563	0.8617	0.8685	1.0908
syntactical	hasHashtag	0.1111	0.0917	0.1166	0.0843	0.0801	0.1274
	hasURL	1.3509	1.1706	1.2355	1.2676	1.3503	1.0556
	isReply	-0.5603	-0.5958	-0.6466	-0.5162	-0.4443	-0.7643
	length	0.0013	-0.0007	0.0003	0.0004	0.0016	-0.0020
semantics	#entities	0.0572	0.0117	0.0620	0.0208	0.0478	-0.0115
	#entities(person)	-0.2613	0.0552	-0.5400	0.0454	0.1088	-0.3932
	#entities(organization)	-0.0952	-0.1767	-0.2257	-0.0409	-0.1636	-0.0297
	#entities(location)	-0.1446	0.0136	-0.1368	-0.1056	-0.0583	-0.1305
	#entities(artifact)	-0.3442	-0.3725	-0.4834	-0.3086	-0.2260	-0.4835
	#entities(species)	-0.2567	-0.9599	-0.8893	-0.4792	-0.1634	-18.8129
	diversity	0.1940	0.2695	0.2776	0.1943	0.1071	0.3867
	sentiment	-0.7968	-0.1761	-0.6297	-0.4727	-0.3227	-0.7411
contextual	social context	-0.002	-0.0068	-0.0020	-0.0057	-0.0034	-0.0055

Table 5: Influence comparison of different features among different topic partitions. There are three splits shown here: popular vs. unpopular topics, global vs. local topics and persistent vs. occasional topics. While the performance measures are based on 5-fold cross-validation, the derived feature weights for the logistic regression model were determined across all topics of a split. The total number of topics is 49. For each topic split, the three features with the highest absolute coefficient are underlined. The extreme negative coefficient for *#entities(species)* and the occasional topic split is an artifact of the small training size: in none of the relevant tweets did this concept type occur.

partitions each and one split with five partitions:

- Popular/unpopular: The topics were split into popular (interesting to many users) and unpopular (interesting to few users) topics. An example of a popular topic is *2022 FIFA soccer* (MB002) - in total we found 24. In contrast, topic *NIST computer security* (MB005) was classified as unpopular (as one of 25 topics).
- Global/local: In this split, we considered the interest for the topic across the globe. The already mentioned topic MB002 is of global interest, since soccer is a highly popular sport in many countries, whereas topic *Cuomo budget cuts* (MB019) is mostly of local interest to users living or working in New York where Andrew Cuomo is the current governor. We found 18 topics to be of global and 31 topics to be of local interest.
- Persistent/occasional: This split is concerned with the interestingness of the topic over time. Some topics persist for a long time, such as MB002 (the FIFA world cup will be played in 2022), whereas other topics are only of short-term interest, e.g. *Keith Olbermann new job* (MB030). We assigned 28 topics to the persistent and 21 topics to the occasional topic partition.
- Topic themes: The topics were classified as belonging to one of five themes, either business, entertainment, sports, politics or technology. While MB002 is a sports topic, MB019 for instance is considered to be a political topic.

Our discussion of the results focuses on two aspects: (i) the difference between the models derived for each of the two partitions, and, (ii) the difference between these models (denoted $M_{splitName}$) and the model derived over all topics ($M_{allTopics}$) in Table 4. The results for the three binary topic splits are shown in Table 5.

Popularity: A comparison of the most important features of $M_{popular}$ and $M_{unpopular}$ shows few differences with the exception of a single feature: sentiment. While sentiment, and in particular a negative sentiment, is the third most important feature in $M_{popular}$, it is ranked eighth in $M_{unpopular}$. We hypothesize that unpopular topics are also partially unpopular because they do not evoke strong emotions in the users. A similar reasoning can be applied when considering the amount of relevant tweets discovered for both topic splits: while on average 67.3 tweets were found to be relevant for popular topics, only 49.9 tweets were found to be relevant for unpopular topics (the average number of relevant tweets across the entire topic set is 58.44).

Global vs. local: This split did not result in models that are significantly different from each other or from $M_{allTopics}$, indicating that—at least for our currently investigated features—a distinction between global and local topics is not useful.

Temporal persistence: The same conclusion can be drawn about the temporal persistence topic split; for both models the same features are of importance which in turn are similar to $M_{allTopics}$. However, it is interesting to see that the performance (regarding all metrics) is clearly higher for the occasional (short-term) topics in comparison to the persistent (long-term) topics. For topics that have a short lifespan recall and precision are notably higher than for the other types of topics.

Topic Themes: The results of the topic split according to the theme of the topic are shown in Table 6. Three topics did not fit in one of the five categories. Since the topic set is split into five partitions, the size of some partitions is extremely small, making it difficult to reach conclusive results. We can, though, detect trends, such as the fact that relevant tweets for business topics are less likely to contain hashtags (negative coefficient), while the opposite holds for entertainment topics (positive coefficient). The

Performance	Measure	business	entertainment	sports	politics	technology
	#topics	6	12	5	21	2
	#samples	4503	9724	4669	17162	1811
	precision	0.4659	0.3691	0.1918	0.3433	0.5109
	recall	0.7904	0.5791	0.1045	0.4456	0.4653
	F-measure	0.5862	0.4508	0.1353	0.3878	0.4870
Feature Category	**Feature**	**business**	**entertainment**	**sports**	**politics**	**technology**
keyword-based	keyword-based	0.2143	0.2069	0.1021	0.1728	0.2075
semantic-based	semantic-based	0.2287	0.2246	0.0858	0.0456	0.0180
	isSemanticallyRelated	1.3821	0.4088	1.0253	1.0689	2.1150
syntactical	hasHashtag	-0.8488	0.5234	0.3752	-0.0403	-0.1503
	hasURL	2.0960	1.1429	1.2785	1.2085	0.4452
	isReply	-0.2738	-0.4784	-0.6747	-0.9130	-0.3912
	length	0.0044	0.0011	0.0050	-0.0009	0.0013
semantics	#entities	-0.2473	-0.1470	0.0853	0.0537	0.1011
	#entities(person)	-1.2929	-0.1161	-0.4852	0.0177	0.1307
	#entities(organization)	-0.0976	0.0865	-0.4259	-0.0673	-0.7318
	#entities(location)	-1.3932	-0.9327	0.3655	-0.1169	0.0875
	#entities(artifact)	-0.4003	-0.1235	-1.0891	-0.2663	-0.3943
	#entities(species)	0.0241	-19.1819	-31.0063	-0.5570	-0.6187
	diversity	0.5277	0.4540	0.3209	0.2037	0.1431
	sentiment	-1.0070	-0.3477	-1.0766	-0.5663	-0.2180
contextual	social context	-0.0067	-0.0086	-0.0047	-0.0041	-0.0155

Table 6: In line with Table 5, this table shows the influence comparison of different features when partitioning the topic set according to five broad topic themes.

semantic similarity has a large impact on all themes but entertainment. Another interesting observation is that sentiment, and in particular negative sentiment, is a prominent feature in $M_{business}$ and in $M_{politics}$ but less so in the other models.

Finally we note that there are also some features which have no impact at all, independent of the topic split employed: the length of the tweet and the social context of the user posting the message. The observation that certain topic splits lead to models that emphasize certain features also offers a natural way forward: if we are able to determine for each topic in advance to which theme or topic characteristic it belongs to, we can select the model that fits the topic best.

6. CONCLUSIONS

In this paper, we have analyzed features that can be used as indicators of a tweet's relevance and interestingness to a given topic. To achieve this, we investigated features along two dimensions: topic-dependent features and topic-independent features. We evaluated the utility of these features with a machine learning approach that allowed us to gain insights into the importance of the different features for the relevance classification.

Our main discoveries about the factors that lead to relevant tweets are the following: (i) The learned models which take advantage of semantics and topic-sensitive features outperform those which do not take the semantics and topic-sensitive features into account. (ii) The length of tweets and the social context of the user posting the message have little impact on the prediction. (iii) The importance of a feature differs depending on the characteristics of the topics. For example, the sentiment-based feature is more important for popular than for unpopular topics and the semantic similarity does not have a significant impact on entertaining topics.

The work presented here is beneficial for search & retrieval of microblogging data and contributes to the foundations of engineering search engines for microposts. In the future, we plan to investigate the social and the contextual features in depth. Moreover, we would like to investigate to what extent personal interests of the users (possibly aggregated from different Social Web platforms) can be utilized as features for personalized retrieval of microposts.

7. REFERENCES

[1] F. Abel, I. Celik, and P. Siehndel. Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter. In *ISWC '11*, Springer, 2011.

[2] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *UIST '10*, ACM, 2010.

[3] J. Chen, R. Nairn, and E. H. Chi. Speak Little and Well: Recommending Conversations in Online Social Streams. In *CHI '11*, ACM, 2011.

[4] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In *WWW '10*, ACM, 2010.

[5] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. In *COLING '10*, Association for Computational Linguistics, 2010.

[6] A. Jadhav, H. Purohit, P. Kapanipathi, P. Ananthram, A. Ranabahu, V. Nguyen, P. N. Mendes, A. G. Smith, M. Cooney, , and A. Sheth. Twitris 2.0 : Semantically Empowered System for Understanding Perceptions From Social Data. In *Semantic Web Challenge*, 2010.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW '10*, ACM, 2010.

[8] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *SIGMOD '10*, ACM, 2010.

[9] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci '11*, 2011.

[10] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10*, ACM, 2010. ACM.

[11] J. Teevan, D. Ramage, and M. R. Morris. #TwitterSearch: a comparison of microblog search and web search. In *WSDM '11*, ACM, 2011.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM '10*, ACM, 2010.

[13] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, ACM, 2001. ACM.