Comparing Tweet Classifications by Authors' Hashtags, Machine Learning, and Human Annotators

Chifumi Nishioka
ZBW – Leibniz Information Centre for Economics, Germany
Kiel University, Germany
Email: chni@informatik.uni-kiel.de

Ansgar Scherp
ZBW – Leibniz Information Centre for Economics, Germany
Kiel University, Germany
Email: asc@informatik.uni-kiel.de

Klaas Dellschaft
WeST – Web Science and Technologies
University of Koblenz-Landau, Germany
Email: klaasd@uni-koblenz.de

***Abstract*—Over the last years, many papers have been published about how to use machine learning for classifying postings on microblogging platforms like Twitter, e. g., in order to assist users to reach tweets that interest them. Typically, the automatic classification results are then evaluated against a gold standard classification which consists of either (i) the hashtags of the tweets' authors, or (ii) manual annotations of independent human annotators. In this paper, we show that there are fundamental differences between these two kinds of gold standard classifications, i. e., human annotators are more likely to classify tweets like other human annotators than like the tweets' authors. Furthermore, we discuss how these differences may influence the evaluation of automatic classifications, like they may be achieved by Latent Dirichlet Allocation (LDA). We argue that researchers who conduct machine learning experiments for tweet classification should pay particular attention to the kind of gold standard they use. One may even argue that hashtags are not appropriate as a gold standard for tweet classification.**

I. Introduction

Twitter is one of the most popular microblogging platforms. Users can post short text messages (tweets) of up to 140 characters long. Furthermore, the authors can annotate their tweets with hashtags. An example tweet is "#Sochi2014 opening ceremony was great!". The hashtag "#Sochi2014" can be seen as a kind of classification for the tweet.

Due to the growth in volume of tweets, it is not trivial to develop methods that help users to access tweets that they are interested in. Recent studies have confirmed that tweet classification could assist users to understand content of tweets and search tweets in the Twitter space [1], [2], [3]. Therefore, tweet classification is vital, in order to organize the massive volume of tweets. Consequently, many authors propose to use machine learning algorithms for automatically classifying tweets (see Section II).

In principle, two main approaches can be identified how to evaluate such automatic classifications: Either they are compared to a gold standard which consists of the authors' hashtags [2], or they are compared to a gold standard which consists of human annotations. For example, Ren et al. [4] used the annotations made by 81 social media experts, who were specifically trained for the task. Also Yang et al. [5] used human annotations as gold standard, but in their experiment, the annotators were only allowed to use terms that are contained in a pre-defined taxonomy.

In this paper, we study in how far there are fundamental differences between these two kinds of gold standard classifications. It is important to understand their differences, because they influence how to interpret the evaluation results of automatic tweet classification approaches. For this purpose, we compare the classification results achieved by authors' hashtags, a machine learning approach, and human annotations. As our dataset, we use tweets and hashtags from the Tweets2011 dataset provided by TREC and prepare ten topics (i. e., ten classification tasks) from the dataset. On these topics, we then apply Latent Dirichlet Allocation (LDA) [6], which can be used for automatically classifying the tweets. Using LDA avoids any bias in our automatic classification towards a specific classification because LDA is an unsupervised approach, which requires no labeled training data. Furthermore, for collecting human annotations, we have conducted an online experiment where 163 participants manually classified the tweets from the ten topics. From the obtained results, we argue that researchers should be aware of which gold standard they use for tweet classification when interpreting results.

The rest of this paper is structured as follows: In Section II, we review the related work with regard to automatically classifying tweets, and how to conduct experiments in which humans manually classify tweets. Then, in Section III, we introduce the three classification approaches hashtag classifier, machine learning classifier, and human classifier in more detail. In Section IV, we describe the preparation of our experiments, i .e., the dataset used and how we collect the tweet classification results using the machine learning approach and human annotators. After that, we investigate and discuss our main research questions, in how far there are differences between the three classification approaches (see Section V), and how the inter-annotator agreement of the human classifiers relates to the agreement between the

human classifiers and the hashtag classifier (see Section VI), before we conclude the paper.

II. Related Work

It already exists quite a lot of work which deals with the problem of classifying short text messages. The two main challenges are the shortness and the sparseness of the texts that should be classified. In other words, short texts are less topic-focused, contain more noise, and they do not provide enough word co-occurrences to compute similarities [7]. In general, one can distinguish between supervised and unsupervised methods for classifying short texts.

Long et al. [1] proposed a semi-supervised classification approach, exploiting transfer learning using Wikipedia as external data. Yang et al. [5] classified tweets based on an inference mechanism for combining texts with additional sources of information. As a result, they assign tweets to nodes in a pre-defined taxonomy. Ren et al. [4] dealt with the hierarchical multi-label classification task for tweets using structural support vector machine (SVM).

However, the supervised methods require a labeled training dataset, which introduces a bias of the classification results towards the training data. In order to avoid such a bias, we use in this paper an unsupervised classification approach. A popular unsupervised approach for short text classification is Latent Dirichlet Allocation (LDA) [6]. LDA represents a document as a probability distribution over the topics and a topic as a probability distribution over the words in a document collection. A document may be an individual tweet or an aggregation of all tweets of a certain user [8]. Hong et al. [8] showed that LDA with aggregated user tweets outperforms LDA with individual tweets. Further examples of classifying aggregated user tweets with LDA are available in [9], [10], [11]. Regarding the content of tweets, Feng et al. [12] showed that the textual content of the tweets is the most important feature for hashtag recommendation. Less important features are the URL links, mentions (i. e., user account names in tweets, like "India's @narendramodi tweets congratulatory wishes to AAP's @ArvindKejriwal"), and hashtag features. However, the combination of all features performed best for hashtag recommendation, where hashtag features slightly contributed to improve recommendation performance.

With regard to how to use human annotators for classifying tweets, there also exist some studies. Paul et al. [13] used Amazon Mechanical Turk (AMT) for classifying tweets by their purpose (i. e., whether it is a question or not). Each tweet was labeled by two AMT workers. Finin et al. [14] used AMT and CrowdFlower[1] for named entity extraction from tweets. In contrast, in this paper, at least ten participants manually classify tweets according to their content. Furthermore, we evaluate the inter-annotator agreement of the participants.

[1]http://www.crowdflower.com/, last access: 08/06/2015

III. Classification Approaches

As stated in the introduction, we compare three classification approaches for microblog postings, namely the *hashtag-classifier*, the *machine-classifier*, and the *human-classifier*.

The **hashtag-classifier** uses the authors' hashtags for assigning tweets to classes. The approach can be divided into two variants: single-label classification and multi-label classification. In the single-label classification, each tweet belongs to only one class. In contrast, in the multi-label classification, each tweet can belong to several classes. Since a tweet can be annotated with multiple hashtags, some studies see the tweet classification problem as a multi-label classification problem [4], [5]. However, we focus on the single-label classification problem, in order to be better comparable with the *machine-classifier* and the *human-classifier*. It means that a tweet that is annotated with #apple and #fruit is in another class than tweets that are only annotated with one of the two hashtags, i. e., only with #apple or #fruit.

The **machine-classifier** classifies tweets based on their textual content. In this paper, we employ Latent Dirichlet Allocation (LDA) [6]. We choose LDA, because it is an unsupervised approach (see Section II). Following the results shown in [8], we train the topic model over aggregated user tweets, i .e., each document represents all tweets generated by a single user. Subsequently, LDA infers for each tweet the probability distribution over the topics from the previously estimated topic model. Finally, we classify the tweets by K-means clustering, which is also an unsupervised clustering algorithm. For this purpose, we represent each tweet by a vector which contains its topic probabilities, as they have been inferred by LDA.

For the **human-classifier**, we asked human annotators to manually classify tweets coming from one of ten different topics as shown in Table I. During the classification, the human annotator was able to see the textual content of the tweets and, if available, the linked web pages (see Section IV-C for more details). Since the tweet classification is expected to highly depend on each human annotator's view, we collected data from 163 participants. Each tweet is classified by at least ten participants.

IV. Preparing the Experiments

In the following, we first describe the Twitter dataset used in this paper. Then, in Section IV-B, we describe the experiment for the *machine-classifier*. Finally, in Section IV-C, we describe how we collected the tweet classifications by the human annotators.

A. Used Twitter Dataset

We use the Tweets2011 dataset provided by TREC[2]. The dataset contains approximately 16 million tweets sampled

[2]http://trec.nist.gov/data/tweets/, last access: 04/30/2015

Table I
TEN MAIN TOPIC/SUBTOPIC COMBINATIONS USED IN OUR EXPERIMENTS, EACH CONSISTING OF 15 TWEETS. THE RIGHT COLUMN SHOWS THE NUMBER OF HUMAN ANNOTATORS.

ID	main topic	subtopics	participants
1	#health	#nutrition, #news	20
2	#apple	#iphone, #mac	18
3	#photography	#nature, #art	15
4	#green	#solar, #eco	14
5	#celebrity	#news, #gossip	15
6	#fashion	#news, #shoes	15
7	#fitness	#health, #exercise	18
8	#humor	#quotes, #funny	15
9	#quote	#love, #life	16
10	#travel	#lp, #tips	17

between Jan. 23 and Feb. 8, 2011. First, we randomly sampled hashtags that occur at least 200 times. We see them as main topics. Then, we identify the hashtags which co-occur with one of the main topics at least four times by four different users, and randomly select two of them as subtopics. We removed domain-specific topics (e. g., "Liverpool", "Photography-HDR" (high dynamic range), "Rockmusic"), religious topics, bots (e. g., "#etsybot"), and very general topics (e. g., "#backintheday"). From the remainder, we randomly select ten of the main topic/subtopic combinations for our experiments (see Table I).

As consequence of this sampling procedure, the tweets of each topic combination can only contain three different hashtags (e. g., main topic: "#apple", subtopics: "#iphone", "#mac"). As we use a single-class classification built from the tweets' hashtags (see Section III), there are five possible classes for the tweets (e. g., {#apple}, {#iphone}, {#mac}, {#apple, #iphone}, {#apple, #mac}). For each of these classes, we randomly select three tweets, thus each topic has 15 tweets. When randomly selecting the tweets, the following constraints have been applied: (i) The tweets have to come from different authors, (ii) links contained in the tweet must not be broken, and (iii) the tweet must not be spam (e. g., bot-generated advertisements). For the *machine-classifier* and *human-classifier*, we remove the hashtags from the tweets, in order to avoid a bias towards the hashtags. But if a hashtag is a part of a sentence (e .g., "watch ten goals at the #WorldCup."), we remove only the hashtag symbol "#" (i. e., "watch ten goals at the WorldCup."). Because it would destroy the sentence structure and perhaps its meaning.

B. Machine-Classifier

In the following, we describe the *machine-classifier* experiments, including how we pre-processed the tweets, and how we set the parameters for LDA and K-means clustering. First, in order to generate the topic model, we discarded tweets produced by users who have less than ten tweets in the Tweets2011 dataset. The remaining tweets were then pre-processed by tokenization, stop-word removal, and stemming. Furthermore, we eliminated words that are used by less than 25 users [15]. The topic model has been trained on the resulting dataset of $1,062,419$ tweets with $13,840$ unique words, coming from $60,947$ users.

For training the topic model, we use the LDA implementation JGibbLDA[3]. Along with Griffiths et al. [16], we set as LDA parameters $\alpha = 1.00$, $\beta = 0.1$. Regarding the number of topics, we used the same value as in the experiments described by Hong et al. [8]. The authors experimented with LDA and found that it performed best at $k = 50$ topics on a dataset with $1,992,758$ tweets, which is of similar size and characteristics to our dataset.

The topic model has then been trained over 2000 iterations. Based on the estimated topic model, we computed the topic probabilities for each of the tweets from our main topic/subtopic combinations (see Table I). For computing the similarities of the vectors, we use Euclidean distance. Please note, the clustering results achieved with Euclidean distance is identical with the results one observes with cosine similarity.

Finally, the topic probabilities have been used for classifying the tweets with the help of K-means clustering. The number of clusters was optimized by Hartigan's index and Average Silhouette [17], following Yang et al. [18]. When the two metrics disagree, we choose the number which is closer to the average number of classes made by human annotators (see Table III).

C. Human-Classifier

In order to collect the classification results for the *human-classifier*, we have conducted an online experiment, where participants manually classified tweets.[4]

1) Participants: In order to obtain the results of the *human-classifier*, we recruited 163 participants (75 female) through mailing lists. As incentive for participation, they obtained information about their classification behavior at the end of the experiment. The last column of Table I shows the number of participants assigned to each topic. The participants are on average 34.14 years old (SD: 10.76) and rated their English skills with an average score of 7.66 (SD: 1.49) on a 10 point Likert scale, where higher is better.

2) Procedure: During the experiment, we asked the participants to classify 15 tweets from one of the ten topic combinations listed in Table I. The topic has been randomly selected. Figure 1 shows a screenshot of the experiment web page. During the experiment, the participants had access to the textual contents of the tweets as well as to screenshots of the webpages that are linked in the tweets. For classifying the tweets, the participants were able to create an arbitrary

[3]http://jgibblda.sourceforge.net/, last access: 04/30/2015

[4]The dataset is available from: http://dx.doi.org/10.7802/82

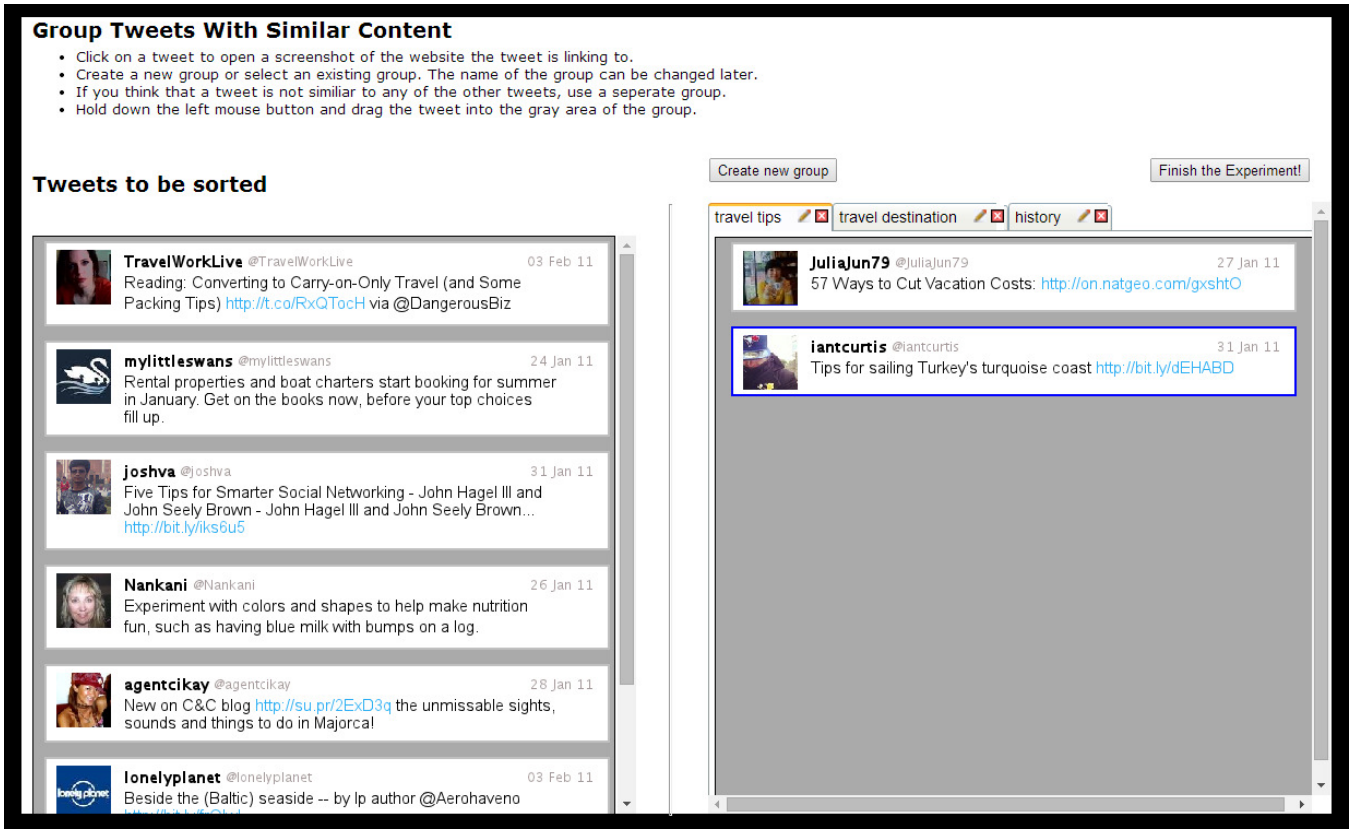

Figure 1. The screenshot of the experimental web page for the *human-classifier*. In the left column, the 15 tweets to be classified are shown. The participants can drag and drop the tweets into the right column to assign the tweets to one of their previously created labeled classes.

number of labeled classes. After assigning all tweets to classes, the participants had to fill out a questionnaire.

V. Agreements between classifiers

In this section, we investigate how strong the agreements between two classifiers are. First, we introduce Cohen's kappa, a measure of the reliability of the agreement, and its interpretation. Subsequently, we show the results with respect to each pair of the three classifiers.

A. Measure

We compute Cohen's kappa κ to measure the reliability of the agreement between two different classifier as Equation 1.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{1}$$

Pr(a) denotes the relative observed agreement between two classifiers. Pr(e) is the hypothetical probability of chance agreement, which is calculated as the probabilities of each classifier randomly assigning each class to tweets, using the observed data. According to Landis et al. [19], $\kappa \leq 0$ is indicating no agreement, $0 < \kappa \leq 0.2$ a slight agreement, $0.2 < \kappa \leq 0.4$ a fair agreement, and $0.8 < \kappa \leq 1$ an almost perfect agreement. If the classifiers agree and disagree completely, $\kappa = 1$ and $\kappa = -1$.

Since the number of classes and the number of tweets in each class might be different depending on the classifiers, we apply Cohen's kappa to so-called match tables [20] (see Table II). The top row of each table indicates how the classifier groups the tweets. We see that the elements are classified into three classes in the left table. If two elements belong to the same class (e. g., a and b), then the corresponding cell value is set to 1. Otherwise, the cell value is set to 0. Based on the match tables in Table II, Cohen's kappa κ is computed as follows: Since six of the ten cells in the two match tables are identical, $Pr(a) = 0.60$. Furthermore, in the left table, two cells are set to 1 and eight are set to 0. In the right table, it is four cells and six cells. Thus, the probability that both tables contain a 1 in the same cell is 0.08. For a cell value of 0, the probability is 0.48. This results in $Pr(e) = 0.08 + 0.48 = 0.56$. Altogether, we obtain a value of $\kappa = 0.18$ (see Equation 1).

In our experiments, we computed the corresponding match tables for each of the topics and classifier separately, i. e., we obtained match tables with 15 rows and columns (one for each tweet). Furthermore, for each topic we ob-

Table II
AN EXAMPLE OF TWO MATCH TABLES (LEFT AND RIGHT) FOR TWEET CLASSIFICATION. THEY ARE PRODUCED BY TWO DIFFERENT CLASSIFIERS. A, B, C, D, E ARE THE CLASSIFIED TWEETS. THE SETS {} INDICATE THE CLASSES.

{a, b}, {c, d}, {e}				
	a	b	c	d
b	1			
c	0	0		
d	0	0	1	
e	0	0	0	0

{a, b, c}, {d, e}				
	a	b	c	d
b	1			
c	1	1		
d	0	0	0	
e	0	0	0	1

Table III
THE NUMBER OF CLASSES AND NUMBER OF TWEETS PER CLASS CREATED BY THE *machine-classifier* AND THE *human-classifier*. STANDARD DEVIATIONS ARE PROVIDED IN PARENTHESES.

	machine-classifier		*human-classifier*	
ID	class	tweets	class	tweets
1	3	5.00 (±4.58)	5.15 (±1.66)	2.91 (±2.24)
2	3	5.00 (±3.00)	3.11 (±1.13)	4.82 (±3.46)
3	4	3.75 (±2.06)	4.20 (±1.86)	3.57 (±3.15)
4	5	3.00 (±1.58)	4.93 (±1.44)	3.04 (±2.06)
5	6	2.50 (±1.97)	4.40 (±1.55)	3.41 (±3.91)
6	4	3.75 (±4.86)	3.73 (±1.53)	4.02 (±4.10)
7	5	3.00 (±2.12)	3.83 (±1.38)	3.91 (±3.53)
8	4	3.75 (±4.86)	4.07 (±1.16)	3.69 (±2.85)
9	6	2.50 (±1.38)	3.75 (±0.86)	4.00 (±2.92)
10	3	5.00 (±6.08)	2.94 (±1.14)	5.10 (±4.48)
M	4.30	3.73	4.01	3.85
SD	1.16	1.00	0.70	0.70

tained one match table for the *hashtag-classifier*, one for the *machine-classifier*, and multiple match tables for the *human-classifier* (one for each participant of the experiment). Given these match tables, we then compute Cohen's kappa pairwise between each participant and one of the other two classifiers. Subsequently, means and standard deviations of Cohen's kappa are calculated over all participants for each topic (see Section V-B).

B. Results and Discussion

In Table III, the number of classes and the average number of tweets per class are shown for the *machine-classifier* and *human-classifier*. For the *hashtag-classifier*, the number of classes and the number of tweets per class are consistently 5 and 3, as described in Section IV-A. These numbers are quite similar to the average number of classes and tweets per class for the *machine-classifier* (4.30 and 3.73) and the *human-classifier* (4.01 and 3.85, cf. Table III).

All in all, it can be noted that there is a high standard deviation for the number of assigned tweets per class for the *machine-classifier*, i. e., it differs a lot between the different topics. In contrast, standard deviations of the *human-classifier* are smaller. Thus, human annotators are more likely to keep the distribution of the number of tweets regardless of the topics.

Comparing hashtag-classifier and machine-classifier: In Table IV, Cohen's kappa is shown for comparing the topics' match tables of the *hashtag-classifier* and the *machine-classifier*. Overall, there is almost no agreement between the classifiers. However, we observe a fair agreement for topic 3 "#photography". A possible reason is that 11 of the 15 tweets in this topic use the hashtags also as a word in their textual content (e .g., "photography", "art"). Thus, the tweet classification by the *machine-classifier* is fairly close to the one made by the *hashtag-classifier* for this topic. Nevertheless, it is all in all difficult for the machine learning approach to reproduce the tweet classifications made by the authors' hashtags.

Table IV
COHEN'S KAPPAS FOR PAIRWISE COMPARISONS BETWEEN THE CLASSIFIERS. FOR COHEN'S KAPPA INVOLVING *human-classifier*, STANDARD DEVIATIONS ARE PROVIDED IN PARENTHESES.

ID	*hashtag* and *machine*	*hashtag* and *human*	*machine* and *human*
1	-0.05	0.12 (±0.08)	0.00 (±0.06)
2	0.02	0.05 (±0.14)	0.05 (±0.19)
3	0.24	0.06 (±0.09)	0.11 (±0.14)
4	0.01	0.11 (±0.14)	0.00 (±0.10)
5	0.00	0.07 (±0.05)	-0.04 (±0.04)
6	0.00	0.15 (±0.13)	0.04 (±0.12)
7	0.04	0.09 (±0.10)	0.05 (±0.10)
8	-0.04	0.17 (±0.13)	0.03 (±0.12)
9	-0.02	0.13 (±0.07)	0.00 (±0.06)
10	0.01	0.10 (±0.08)	0.45 (±0.25)
M	0.02	0.10	0.07
SD	0.08	0.10	0.12

Comparing hashtag-classifier and human-classifier: Table IV also contains the Cohen's kappa values for comparing the *hashtag-classifier* to each of the *human-classifiers*. For all the topics, we consistently observe a slight agreement between the two classifiers.

Comparing machine-classifier and human-classifier: Finally, Table IV also contains the comparison between the *machine-classifier* and each of the *human-classifiers*. Overall, for most of the topics, there is almost no agreement between the classifiers. However, the classifiers have a moderate agreement for topic 10 "#travel", whose subtopics are "#lp", referring to Lonely Planet, a travel publisher, and "#tips". While the tweets which only contain the hashtag "#tips" are apparently about different topics, almost all other tweets are about travel because they either contain "#travel" or "#lp" as a hashtag. Thus, the different classes are clearly separated for this topic, leading to almost the same number of classes for the *machine-classifier* and *human-classifier* (cf. Table III). Furthermore, a high standard deviation in the number of tweets per class can be observed in Table III, which indicates that there is a dominant class for this topic.

VI. AGREEMENT AMONG HUMAN-CLASSIFIERS

In this section, we analyze the inter-annotator agreement between the participants of the *human-classifier* experiment. Additionally, we investigate whether the *huma-classifier*

is more similar to the *hashtag-classifier* or the *machine-classifier*.

A. Measures

For measuring the inter-annotator agreement of the human annotators, we use Fleiss' kappa. Additionally, we compute purity, conditional entropy, and Normalized Mutual Information (NMI). For these measures, we use the classifications of the *hashtag-classifier* and the *machine-classifier* as gold standards.

Fleiss' kappa: We use Fleiss' kappa κ to assess the inter-annotator agreement of the human annotators for the different topics. In contrast to Cohen's kappa, which measures the agreement between two classifiers, Fleiss' kappa can be used for measuring the agreement between more than two classifiers [21]. Fleiss' kappa is computed by Equation 2.

$$\kappa = \frac{\overline{Pr(a)} - \overline{Pr(e)}}{1 - \overline{Pr(e)}} \tag{2}$$

$\overline{Pr(a)}$ denotes the agreement between more than two classifiers, which are in our case the different human annotators. $\overline{Pr(e)}$ is the probability of chance agreement between the classifiers. Fleiss' kappa is interpreted in the same way as Cohen's kappa.

Purity: Purity measures the overall precision of the elements assigned to clusters [22]. The purity of a cluster $c_i \in C$, where $C = \{c_1, \ldots, c_k\}$ is defined by Equation 3.

$$Purity(c_i) = \frac{1}{|c_i|} \cdot max_h |c_i \cap a_h| \tag{3}$$

a_h denotes a cluster from the gold standard, which are in our case the *hashtag-classifier* or the *machine-classifier*. Purity simply measures how many tweets from cluster c_i are contained in the most similar cluster from the gold standard, and puts this into relation to the overall size of cluster c_i. Given this purity for a single cluster c_i, the purity of the entire clustering is computed by Equation 4.

$$Purity(C) = \sum_{i=1}^{k} \frac{|c_i|}{N} \cdot Purity(c_i) \tag{4}$$

N denotes the number of elements in the dataset and k the number of clusters in C that are compared to the gold standard. In our case, $N = 15$ because for each topic 15 tweets have been classified. $Purity$ results in a value between 0 and 1, where a value closer to 1 implies a better agreement between the tested clustering and the gold standard.

Conditional Entropy: Conditional entropy measures how the elements are distributed within the cluster c_i given the various classes a_h from the gold standard [22]. The conditional entropy for a certain cluster c_i is defined by Equation 5.

$$H(c_i|A) = -\sum_{h=1}^{l} P(a_h|c_i) \cdot \log P(a_h|c_i), \tag{5}$$

l denotes the number of clusters contained in the gold standard. The conditional entropy of the whole clustering C is then defined as the sum of the individual cluster entropies weighted according to the cluster size (see Equation 6).

$$H(C|A) = \sum_{i=1}^{k} \frac{|c_i|}{N} \cdot H(c_i|A) \tag{6}$$

The conditional entropy results in values between 0 and 1. A value closer to 0 shows a better agreement between clustering C and the gold standard, because it implies that the elements from the classes a_h are not distributed over different clusters c_i.

Normalized Mutual Information (NMI): A high purity is easy to achieve if the number of clusters is large, i. e., when there are only few elements per cluster. In contrast, the conditional entropy is negatively influenced if there is a larger difference between the number of classes in C and in A. To overcome these drawbacks, we use Normalized Mutual Information (NMI) [23] (see Equation 7). It represents the mutual agreement between two classifiers.

$$NMI = \frac{2 \cdot I(C;A)}{H(C) + H(A)}, \tag{7}$$

$I(C;A)$ denotes the mutual information of the two clusterings C and A, as defined in Equation 8. $H(C)$ and $H(A)$ denotes the entropy of the marginal distributions in the two clusterings, as defined in Equation 9.

$$I(C;A) = \sum_{i=1}^{k} \sum_{h=1}^{l} \frac{|c_i \cap a_h|}{N} \cdot \log \frac{|c_i \cap a_h|}{|c_i| \cdot |a_h|} \tag{8}$$

$$H(C) = -\sum_{i=1}^{k} \frac{|c_i|}{N} \log \frac{|c_i|}{N} \tag{9}$$

NMI results in values between 0 and 1, where a value closer to 1 implies a better agreement between the clustering C and the gold standard A.

B. Results

Fleiss' kappa: Table V shows the inter-annotator agreement between the human annotators, as it is measured by Fleiss' kappa. As one can see, the results highly depend on the topic. However, for all topics, the inter-annotator agreement of the human annotators is consistently higher than the agreement between the human annotators and the *hashtag-classifier* (see the middle column in Table IV) and/or the *machine-classifier* (see the right column in Table IV). Thus, human annotators are more likely to classify tweets like

Table VI
PURITY, CONDITIONAL ENTROPY, AND NORMALIZED MUTUAL INFORMATION (NMI) OF *human-classifier* USING *hashtag-classifier* OR *machine-classifier* AS GOLD STANDARD. STANDARD DEVIATIONS ARE PROVIDED IN PARENTHESES.

	(a) *hashtag-classifier*			(b) *machine-classifier*		
ID	(a1) Purity	(a2) Entropy	(a3) NMI	(b1) Purity	(b2) Entropy	(b3) NMI
1	0.66 (±0.11)	0.38 (±0.06)	0.51 (±0.11)	0.49 (±0.14)	0.71 (±0.10)	0.24 (±0.09)
2	0.72 (±0.17)	0.41 (±0.23)	0.30 (±0.19)	0.65 (±0.20)	0.57 (±0.28)	0.22 (±0.19)
3	0.69 (±0.15)	0.39 (±0.16)	0.38 (±0.18)	0.65 (±0.14)	0.45 (±0.17)	0.35 (±0.18)
4	0.64 (±0.10)	0.43 (±0.15)	0.48 (±0.15)	0.58 (±0.09)	0.55 (±0.11)	0.38 (±0.11)
5	0.77 (±0.14)	0.25 (±0.12)	0.44 (±0.06)	0.77 (±0.09)	0.36 (±0.05)	0.34 (±0.09)
6	0.84 (±0.11)	0.21 (±0.11)	0.44 (±0.20)	0.71 (±0.16)	0.50 (±0.21)	0.24 (±0.13)
7	0.75 (±0.14)	0.34 (±0.17)	0.40 (±0.16)	0.74 (±0.15)	0.41 (±0.16)	0.36 (±0.15)
8	0.75 (±0.10)	0.34 (±0.12)	0.48 (±0.17)	0.60 (±0.14)	0.62 (±0.17)	0.28 (±0.09)
9	0.76 (±0.09)	0.34 (±0.11)	0.46 (±0.06)	0.68 (±0.12)	0.45 (±0.15)	0.36 (±0.08)
10	0.86 (±0.14)	0.21 (±0.16)	0.39 (±0.16)	0.85 (±0.10)	0.34 (±0.18)	0.46 (±0.22)
M	0.74	0.33	0.43	0.67	0.50	0.32
SD	0.13	0.14	0.14	0.13	0.16	0.13

Table V
FLEISS' KAPPA FOR MEASURING THE AGREEMENTS AMONG HUMAN ANNOTATORS ON HOW THEY CLASSIFY TWEETS.

ID	Fleiss' kappa
1	0.17
2	0.10
3	0.13
4	0.16
5	0.53
6	0.20
7	0.14
8	0.31
9	0.33
10	0.38
M	0.25
SD	0.14

other human annotators than like the tweets authors and/or a machine learning approach.

Looking into the results for each topic, the best Fleiss' kappa is achieved for topic 5 "#celebrity", indicating a moderate agreement between the human annotators. This topic includes two tweets that are very different from the other tweets in the topic. As a result, 11 of the 15 human annotators created two specific classes "economy" and "science" for these two tweets, and put none of the other tweets into these classes. This explains the high inter-annotator agreement for topic 5.

Purity: Table VI contains the purity values for comparing the *human-classifier* to the gold standard defined by the *hashtag-classifier* (column (a1)) and the *machine-classifier* (column (b1)). The results show that the *human-classifier* is more similar to the *hashtag-classifier* than to the *machine-classifier*. In order to determine whether the differences between the purity values are significant, we run a mixed ANOVA test with one between subject factor (i. e., topics) and one within subject factor (i. e., the classifiers). The analysis reveals that all factors including topics ($F(0.03, 0.23) = 7.43$, $p = .00$), classifiers ($F(0.00, 0.43) = 94.36$, $p = .00$), and the interaction of the two factors ($F(0.00, 0.23) = 6.81$, $p = .00$) significantly differ. Thus, the purity values are significantly better when the *hashtag-classifier* is used as the gold standard. Details are omitted here for reasons of brevity.

Conditional Entropy: Table VI also contains the conditional entropy values for comparing the *human-classifier* to the *hashtag-classifier* (column (a2)) and the *machine-classifier* (column (b2)). Since lower entropy values indicate a better agreement between two clusterings, also these results show that the *human-classifier* is more similar to the *hashtag-classifier* than to the *machine-classifier*. Again, a mixed ANOVA test shows that the differences are significant for all factors, i. e., for topics ($F(0.04, 0.27) = 6.47$, $p = .00$), classifiers ($F(0.01, 2.18) = 229.24$, $p = .00$), and the interaction of the two factors ($F(0.00, 0.08) = 8.24$, $p = .00$).

Normalized Mutual Information (NMI): The NMI values for comparing the *human-classifier* to the *hashtag-classifier* and the *machine-classifier* are contained in the columns (a3) and (b3) in Table VI. Also the NMI shows that the *human-classifier* is more similar to the *hashtag-classifier*. A mixed ANOVA test shows that the differences are significant for all factors, i. e., for topics ($F(0.03, 0.08) = 2.33$, $p = .02$), classifiers ($F(0.01, 0.89) = 101.39$, $p = .00$), and the interaction of the two factors ($F(0.01, 0.09) = 9.71$, $p = .00$).

C. Discussion

The results of our experiment show that the *human-classifier* is consistently more similar to the *hashtag-classifier* than to the *machine-classifier*, independent of the used measure and topic. Nevertheless, we also showed that this agreement between the *human-classifier* and the *hashtag-classifier* is not as good as the inter-annotator agreement between the different human annotators (cf. Table IV and V).

Thus, any experiment on tweets classification should be handled with care regarding the gold standard used. The

researchers should be aware in their discussion of evaluation results, whether the authors' hashtags or human annotations are used as a gold standard.

One might even argue that in general it is more reasonable to use human annotations as a gold standard because the authors' hashtags are created in a highly distributed fashion by many different individuals, thus resulting in tweet annotations based on varying criteria. At least, researchers who use the *hashtag-classifier* as their gold standard, like [2], could explicitly reflect on that difference in their discussions. In this line, we contribute to the recent discussion on how to conduct evaluations in social media research [24], where a gold standard is typically not easily available or absent at all.

VII. Conclusions

In this paper, we compared three approaches for classifying tweets, namely using authors' hashtags, machine learning (LDA), and human annotators. In our experiments, there has been no agreement between the clustering achieved with LDA and the clustering by authors' hashtags. This shows that it is quite hard to reproduce the clustering given by authors' hashtags with a state-of-the-art unsupervised machine learning approach, like LDA, that only gets the textual content of the tweets as its input data.

On the other hand, there has been a slight agreement between the classifications by LDA and by the human annotators. This shows that it is not impossible to extract meaningful classifications with LDA, even though the classifications differ from the authors' hashtags.

This difference between the evaluation results, depending on whether we use authors' hashtags or human annotators as a gold standard, seems to be caused by a fundamental difference between the two gold standards. This fundamental difference is indicated by our observation that the inter-annotator agreement between the human annotators is much higher than the agreement between the human annotators and the authors' hashtags.

Therefore, we argue that researchers should reflect in their discussions of tweet classification results, whether they are compared to a gold standard consisting of the authors' hashtags or to human annotations. As the authors' hashtags are created by many different users and thus come from various different contexts and cultures, one might even argue that in general it is more reasonable to use a gold standard made by human annotators for tweet classification tasks.

Acknowledgments: We thank Jochen Hunz for developing the online web experiment and Isabella Peters for constructive suggestions to improve the human labeling experiment.

References

[1] G. Long, L. Chen, X. Zhu, and C. Zhang, "TCSST: Transfer classification of short & sparse text using external data," in *CIKM*. ACM, 2012, pp. 764–772.

[2] K. Nishida, T. Hoshide, and K. Fujimura, "Improving tweet stream classification by detecting changes in word probability," in *SIGIR*. ACM, 2012, pp. 971–980.

[3] S. Zhang, X. Jin, D. Shen, B. Cao, X. Ding, and X. Zhang, "Short text classification by detecting information path," in *CIKM*. ACM, 2013, pp. 727–732.

[4] Z. Ren, M.-H. Peetz, S. Liang, W. van Dolen, and M. de Rijke, "Hierarchical multi-label classification of social text streams," in *SIGIR*. ACM, 2014, pp. 213–222.

[5] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta, "Large-scale high-precision topic modeling on Twitter," in *KDD*. ACM, 2014, pp. 1907–1916.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, 2003.

[7] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *WWW*. ACM, 2008, pp. 91–100.

[8] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *SOMA*. ACM, 2010, pp. 80–88.

[9] D. Quercia, H. Askham, and J. Crowcroft, "TweetLDA: Supervised topic classification and link prediction in Twitter," in *WebSci*. ACM, 2012, pp. 247–250.

[10] R. Li, W. He, Z. Wu, J. Hu, and Y. Liu, "Modeling user's temporal dynamic profile in micro-blogging using clustering method," in *ISSDM*. IEEE, 2012, pp. 808–812.

[11] H. Koga and T. Taniguchi, "Developing a user recommendation engine on Twitter using estimated latent topics," in *ISSDM*. Springer Berlin Heidelberg, 2012, pp. 461–470.

[12] W. Feng and J. Wang, "We can learn your #hashtags: Connecting tweets to explicit topics," in *ICDE*. IEEE, 2014, pp. 856–867.

[13] S. A. Paul, L. Hong, and E. H. Chi, "What is a question? crowdsourcing tweet categorization," in *CHI 2011 workshop on Crowdsourcing and Human Computation*. ACM, 2011.

[14] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in Twitter data with crowdsourcing," in *CSLDAMT*. Association for Computational Linguistics, 2010, pp. 80–88.

[15] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*. ACM, 2006, pp. 113–120.

[16] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[17] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, Inc., 2005.

[18] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *WSDM*. ACM, 2011, pp. 177–186.

[19] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[20] W.-T. Fu and W. Dong, "Collaborative indexing and knowledge exploration: A social learning model," *IEEE Intelligent Systems*, vol. 27, no. 1, pp. 39–46, 2012.

[21] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability." *Educational and psychological measurement*, 1973.

[22] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," *IEEE TKDE*, vol. 20, no. 9, pp. 1217–1229, 2008.

[23] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel K-means: Spectral clustering and normalized cuts," in *KDD*. ACM, 2004, pp. 551–556.

[24] R. Zafarani and H. Liu, "Evaluation without ground truth in social media research," *Commun. ACM*, vol. 58, no. 6, pp. 54–60, 2015.

APPENDIX: QUALITATIVE ANALYSIS

For the *human-classifier*, we have manually investigated how the human annotators classified the tweets by looking into the labels of the classes. Our analysis revealed that some participants classified tweets by their subjective views rather than the topics of the tweets. For instance, several participants created classes like "interesting" and "not interesting", and assign tweets to either of them. In addition, some participants created classes by purposes of tweets, like "advertisement".

In order to further assess this qualitative analysis, we cluster annotators showing similar grouping behavior. First, we converted the match tables shown in Table II into numeric vectors whose elements are 0 or 1, and which correspond to the cell values of the match table. Subsequently, we run a K-means clustering on the vectors. We optimized the number of clusters by using Hartigan's index and Average Silhouette [17]. When the two metrics disagree, we chose the smaller number of clusters. Table VII shows the number of clusters resulting from this procedure with respect to each topic. Overall, the number of clusters lies between 2 and 4 per topic, i. e., there exist two to four different kinds of how the human annotators classified the tweets of a topic.

Subsequently, we manually looked into each of the identified clusters for patterns regarding how participants classify tweets. We observe that human annotators, who produce the same number of classes, i. e., belong to the same cluster, show a similar grouping strategy. Thus, the number of clusters seems to discriminate the different classification behaviours of the human annotators.

Table VII
CLUSTERING OF *human-classifiers*. EACH CLUSTER REPRESENTS A SET OF ANNOTATORS WITH SIMILAR BEHAVIOR OF GROUPING THE TWEETS.

ID	# of clusters
1	2
2	2
3	2
4	4
5	3
6	3
7	2
8	3
9	3
10	3
M	2.70
SD	0.67

One hypothesis with regard to how human annotators classify tweets would be that they are influenced by textual similarities and/or word occurrences in the tweets. In order to further investigate this hypothesis, we also compute the cosine similarities between the tweets that belong to the same class. However, the cosine similarities between tweets from the same class are very low, i. e., on average 0.06. Applying lemmatisation and stopword removal even led to a further decrease of the cosine similarities.

The tweets used in our experiments were selected such that they cover different topics. For some topics the tweets might be easier to classify than for others (e .g., the topic 5 on "#celebrity" has the highest agreement among human classifiers, see Table V). Thus, we investigated whether the understandability of the different topics may have some influence on the experimental results. To this end, we asked the participants in the final questionnaire to assess the understandability of the tweets on a 10-point Likert scale, where higher values indicate higher understandability. On average, the participants evaluate the understandability of the tweets of one topic with 7.37 (SD: 1.93). Statistical tests reveal no significant difference between topics (results omitted for brevity). Thus, one can exclude that some topics were more difficult to understand than others.