Coling 2010: Poster Volume, pages 241–249, Beijing, August 2010 Enhanced Sentiment Learning Using Twitter Hashtags and Smileys Dmitry Davidov* 1 Oren Tsur* 2 1 ICNC / 2 Institute of Computer Science The Hebrew University {oren,arir}@cs.huji.ac.il Ari Rappoport 2 Abstract Automated identification of diverse sentiment types can be beneficial for many NLP systems such as review summarization and public media analysis.Under multi-class classification we attempt to assign a single label (51 labels in case of hashtags and 16 labels in case of smileys) to each of vectors in the test set. Table 2 shows the performance of our framework for these tasks. Results are significantly above the random baseline and definitely nontrivial considering the equal class sizes in the test set. The relatively low performance of hashtags can be explained by ambiguity of the hashtags and some overlap of sentiments. Examination of classified sentences reveals that many of them can be reasonably assigned to more than one of the available hashtags or smileys. Thus a tweet "I'm reading stuff that I DON'T understand again! hahaha...wth am I doing" may reasonably match tags #sarcasm, #damn, #haha, #lol, #humor, #angry etc. Close examination of the incorrectly classified examples also reveals that substantial amount of tweets utilize hashtags to explicitly indicate the specific hashtagged sentiment, in these cases that no sentiment value could be perceived by readers unless indicated explicitly, e.g. "De Blob game review posted on our blog. #fun". Obviously, our framework fails to process such cases and captures noise since no sentiment data is present in the processed text labeled with a specific sentiment label. Binary classification. In the binary classification experiments, we classified a sentence as either appropriate for a particular tag or as not bearHashtags Avg #hate #jealous #cute #outrageous Pn+W-M-Pt- 0.57 0.6 0.55 0.63 0.53 Pn+W+M-Pt- 0.64 0.64 0.67 0.66 0.6 Pn+W+M+Pt- 0.69 0.66 0.67 0.69 0.64 Pn-W-M-Pt+ 0.73 0.75 0.7 0.69 0.69 FULL 0.8 0.83 0.76 0.71 0.78 Smileys Avg :) ;) X(: d Pn+W-M-Pt- 0.64 0.66 0.67 0.56 0.65 Pn+W+M-Pt- 0.7 0.73 0.72 0.64 0.69 Pn+W+M+Pt- 0.7 0.74 0.75 0.66 0.69 Pn-W-M-Pt+ 0.75 0.78 0.75 0.68 0.72 FULL 0.86 0.87 0.9 0.74 0.81 Table 3: Binary classification results for smileys and hashtags.#ihatequotes". Note that in the first example the hashtagged words are a grammatical part of the sentence (it becomes meaningless without them) while #ihateqoutes of the second example is a mere sentiment label and not part of the sentence. Also note that hashtags can be composed of multiple words (with no spaces). 5 Identification of proper English words was based on an available WN-based English dictionary 244 Category # of tags % agreement Strong sentiment 52 87 Likely sentiment 70 66 Context-dependent 110 61 Focused 45 75 No sentiment 3564 99 Table 1: Annotation results (2 judges) for the 3852 most frequent tweeter tags. The second column displays the average number of tags, and the last column shows % of tags annotated similarly by two judges. During preprocessing, we have replaced URL links, hashtags and references by URL/REF/TAG meta-words. This substitution obviously had some effect on the pattern recognition phase (see Section 3.1.2), however, our algorithm is robust enough to overcome this distortion. 4.2 Hashtag-based sentiment labels The Twitter dataset contains above 2.5 million different user-defined hashtags. Many tweets include more than a single tag and 3852 "frequent" tags appear in more than 1000 different tweets.If a human subject failed to provide the intended "correct" answer to at least two of the control set questions we reject him/her from the calculation. In our evaluation the algorithm is considered to be correct if one of the tags selected by a human judge was also selected by the algorithm. Table 4 shows results for human judgement classification. The agreement score for this task was ? = 0.41 (we consider agreement when at least one of two selected items are shared). Table 4 shows that the majority of tags selected by humans matched those selected by the algorithm. Precision of smiley tags is substantially Setup % Correct % No sentiment Control Smileys 84% 6% 92% Hashtags 77% 10% 90% Table 4: Results of human evaluation. The second column indicates percentage of sentences where judges find no appropriate tags from the list. The third column shows performance on the control set. Hashtags #happy #sad #crazy # bored #sad 0.67 – – – #crazy 0.67 0.25 – – #bored 0.05 0.42 0.35 – #fun 1.21 0.06 1.17 0.43 Smileys :) ;) : (X(;) 3.35 – – – : (3.12 0.53 – – X(1.74 0.47 2.18 – : S 1.74 0.42 1.4 0.15 Table 5: Percentage of co-appearance of tags in tweeter corpus. higher than of hashtag labels, due to the lesser number of possible smileys and the lesser ambiguity of smileys in comparison to hashtags. 5.3 Exploration of feature dependencies Our algorithm assigns a single sentiment type for each tweet. However, as discussed above, some sentiment types overlap (e.g., #awesome and #amazing). Many sentences may express several types of sentiment (e.g., #fun and #scary in "Oh My God http://goo.gl/fb/K2N5z #entertainment #fun #pictures #photography #scary #teaparty").We selected 50 hashtags annotated "1" or "2" by both judges. For each of these tags we automatically sampled 1000 tweets resulting in 50000 labeled tweets. We avoided sampling tweets which include more than one of the sampled hashtags. As a no-sentiment dataset we randomly sampled 10000 tweets with no hashtags/smileys from the whole dataset assuming that such a random sample is unlikely to contain a significant amount of sentiment sentences. 4.3 Smiley-based sentiment labels While there exist many "official" lists of possible ASCII smileys, most of these smileys are infrequent or not commonly accepted and used as sentiment indicators by online communities.If there are no matching vectors found for v, we assigned the default "no sentiment" label since there is significantly more non-sentiment sentences than sentiment sentences in Twitter. 4 Twitter dataset and sentiment tags In our experiments we used an extensive Twitter data collection as training and testing sets. In our training sets we utilize sentiment hashtags and smileys as classification labels. Below we describe this dataset in detail. 4.1 Twitter dataset We have used a Twitter dataset generously provided to us by Brendan O'Connor. This dataset includes over 475 million tweets comprising roughly 15% of all public, non-"low quality" tweets created from May 2009 to Jan 2010.Apart of simple text, tweets may contain URL addresses, references to other Twitter users (appear as @) or a content tags (also called hashtags) assigned by the tweeter (#) which we use as labels for our supervised classification framework.By utilizing 50 Twitter tags and 15 smileys as sentiment labels, this framework avoids the need for labor intensive manual annotation, allowing identification and classification of diverse sentiment types of short texts.5.2 Evaluation with human judges In the second set of experiments we evaluated our framework on a test set of unseen and untagged tweets (thus tweets that were not part of the train6Note that this is a useful application in itself, as a filter that extracts sentiment sentences from a corpus for further focused study/processing.We applied our framework with its FULL setting, learning the sentiment tags from the training set for hashtags and smileys (separately) and executed the framework on the reduced Tweeter dataset (without untagged data) allowing it to identify at least five sentences for each sentiment class. While the majority of existing sentiment extraction systems focus on polarity identification (e.g., positive vs. negative reviews) or extraction of a handful of pre-specified mood labels, there are many useful and relatively unexplored sentiment types.It was suggested that sentiment words may have different senses (Esuli and Sebastiani, 2006; Andreevskaia and Bergler, 2006; Wiebe and Mihalcea, 2006), thus word sense disambiguation can improve sentiment analysis systems (Akkaya et al., 2009).Fully automated evaluation allowed us to test the performance of our algorithm under several different feature settings: Pn+W-M-Pt-, Pn+W+M-Pt-, Pn+W+M+Pt-, Pn-W-M-Pt+ and FULL, where +/- stands for utilization/omission of the following feature types: Pn:punctuation, W:Word, M:ngrams (M stands for 'multi'), Pt:patterns.Mihalcea and Liu (2006) derive lists of words and phrases with happiness factor from a corpus of blog posts, where each post is annotated by the blogger with a mood label.This pattern based framework was proven efficient for sarcasm detection in (Tsur et al., 2010; 2Note that the FH and FC bounds allow overlap between some HFWs and CWs.Two human judges manually annotated these frequent tags into five different categories: 1 - strong sentiment (e.g #sucks in the example above), 2 - most likely sentiment (e.g., #notcute), 3 - contextdependent sentiment (e.g., #shoutsout), 4 - focused sentiment (e.g., #tmobilesucks where the target of the sentiment is part of the hashtag), and 5 - no sentiment (e.g. #obama).2 Related work Sentiment analysis tasks typically combine two different tasks: (1) Identifying sentiment expressions, and (2) determining the polarity (sometimes called valence) of the expressed sentiment. Others combine additional feature types for this decision (Yu and Hatzivassiloglou, 2003; Kim and Hovy, 2004; Wilson et al., 2005; Bloom et al., 2007; McDonald et al., 2007; Titov and McDonald, 2008a; Melville et al., 2009).Another line of works aims at identifying a broader range of sentiment classes expressing various emotions such as happiness, sadness, boredom, fear, and gratitude, regardless (or in addition to) positive or negative evaluations.Comparing the contributed performance of different feature types we can see that punctuation, word and pattern features, each provide a substantial boost for classification quality while we observe only a marginal boost when adding n-grams as classification features.We evaluate the contribution of different feature types for sentiment classification and show that our framework successfully identifies sentiment types of untagged sentences.Moreover, the assigned tag applies to the whole blog post while a finer grained sentiment extraction is needed (McDonald et al., 2007).We utilize 50 Twitter tags and 15 smileys as sentiment labels which allow us to build a classifier for dozens of sentiment types for short textual sentences.Balog et al. (2006) use the mood annotation of blog posts coupled with news data in order to discover the events that drive the dominant moods expressed in blogs. While most of the works on sentiment analysis focus on full text, some works address sentiment analysis in the phrasal and sentence level, see (Yu and Hatzivassiloglou, 2003; Wilson et al., 2005; McDonald et al., 2007; Titov and McDonald, 2008a; Titov and McDonald, 2008b; Wilson et al., 2009; Tsur et al., 2010) among others.Since the patterns we use are relatively long, exact matches are uncommon, and taking advantage of partial matches allows us to significantly reduce the sparsity of the feature vectors.3.1.3 Efficiency of feature selection Since we avoid selection of textual features which have a training set frequency below 0.5%, we perform feature selection incrementally, on each stage using the frequencies of the features obtained during the previous stages.Thus first we estimate the frequencies of single words in the training set, then we only consider creation of n-grams from single words with sufficient frequency, finally we only consider patterns composed from sufficiently frequent words and ngrams.All these features were normalized by dividing them by the (maximal observed value times averaged maximal value of the other feature groups), thus the maximal weight of each of these features is equal to the averaged weight of a single pattern/word/n-gram feature.5 Evaluation and Results The purpose of our evaluation was to learn how well our framework can identify and distinguish between sentiment types defined by tags or smileys and to test if our framework can be successfully used to identify sentiment types in new untagged sentences.For each of the 50 (15) labels for hashtags (smileys) we have performed a binary classification when providing as training/test sets only positive examples of the specific sentiment label together with non-sentiment examples.We can see that even for binary classification settings, classification of smiley-labeled sentences is a substantially easier task compared to classification of hashtag-labeled tweets.To ensure the quality of assignments, we added to each test five manually selected, clearly sentiment bearing, assignment tasks from the tagged Twitter sentences used in the training set.Automated identification of diverse sentiment types can be beneficial for many NLP systems such as review summarization systems, dialogue systems and public media analysis systems.Several works (Wiebe, 2000; Turney, 2002; Riloff, 2003; Whitelaw et al., 2005) use lexical resources and decide whether a sentence expresses a sentiment by the presence of lexical items (sentiment words).While (Mishne, 2005) classifies a blog entry (post), (Mihalcea and Liu, 2006) assign a happiness factor to specific words and expressions.Davidov et al. (2010) analyze the use of the #sarcasm hashtag and its contribution to automatic recognition of sarcastic tweets.3.1.1 Word-based and n-gram-based features Each word appearing in a sentence serves as a binary feature with weight equal to the inverted count of this word in the Twitter corpus.For each feature vector V in the test set, we compute the Euclidean distance to each of the matching vectors in the training set, where matching vectors are defined as ones which share at least one pattern/n-gram/word feature with v. Let ti , i = 1 .We used the Amazon Mechanical Turk (AMT) service in order to obtain a list of the most commonly used and unambiguous ASCII smileys.From the obtained list of smileys we selected a subset of 15 smileys which were (1) provided by at least three human subjects, (2) described by at least two human subject using the same single-word description, and (3) appear at least 1000 times in our Twitter dataset.We then sampled 1000 tweets for each of these smileys, using these smileys as sentiment tags in the sentiment classification framework described in the previous section.Obviously the results are much better than those of multi-class classification, and the observed0.8 precision confirms the usefulness of the proposed framework for sentiment classification of a variety of different sentiment types.We also explore dependencies and overlap between different sentiment types represented by smileys and Twitter hashtags.1 Introduction A huge amount of social media including news, forums, product reviews and blogs contain numerous sentiment-based sentences. In our study we use four different feature types (punctuation, words, n-grams and patterns) for sentiment classification and evaluate the contribution of each feature type for this task.We also explore the dependencies and overlap between different sentiment types represented by smileys and Twitter tags.Mishne (2005) used an ontology of over 100 moods assigned to blog posts to classify blog texts according to moods.Strapparava and Mihalcea (2008) classify blog posts and news headlines to six sentiment categories.ASCII smileys and other punctuation sequences containing two or more consecutive punctuation symbols were used as single-word features.3As with word and n-gram features, the maximal feature weight of a pattern p is defined as the inverse count of a pattern in the complete Twitter corpus.3.2 Classification algorithm In order to assign a sentiment label to new examples in the test set we use a k-nearest neighbors (kNN)-like strategy.We asked each of ten AMT human subjects to provide at least 6 commonly used ASCII mood-indicating smileys together with one or more single-word descriptions of the smiley-related mood state.To avoid utilization of labels as strong features in the test set, we removed all instances of involved label hashtags/smileys from the tweets used as the test set.Avg column shows averaged harmonic f-score for 10- fold cross validation over all 50(15) sentiment hashtags (smileys).In order to make the evaluation harsher, we removed all tweets containing at least one of the relevant classification hashtags (or smileys).One tag from this list was provided by our algorithm, 8 other tags were sampled from the remaining 49 (14) available sentiment tags, and the tenth tag is from the list of frequent non-sentiment tags (e.g. travel or obama).Sentiment is defined as "a personal belief or judgment that * * Both authors equally contributed to this paper. is not founded on proof or certainty"1 .In some other cases obtaining sentiment value can greatly enhance information extraction tasks like review summarization.??0