Opinion Mining on YouTube

Aliaksei Severyn1, Alessandro Moschitti3,1, Olga Uryupina1, Barbara Plank2, Katja Filippova4
1DISI - University of Trento, 2CLT - University of Copenhagen, 3Qatar Computing Research Institute, 4Google Inc.
severyn@disi.unitn.it, amoschitti@qf.org.qa, uryupina@gmail.com, bplank@cst.dk, katjaf@google.com

Abstract

This paper defines a systematic approach to Opinion Mining (OM) on YouTube comments by (i) modeling classifiers for predicting the opinion polarity and the type of comment and (ii) proposing robust shallow syntactic structures for improving model adaptability. We rely on the tree kernel technology to automatically extract and learn features with better generalization power than bag-of-words. An extensive empirical evaluation on our manually annotated YouTube comments corpus shows a high classification accuracy and highlights the benefits of structural models in a cross-domain setting.

1 Introduction

Social media such as Twitter, Facebook or YouTube contain rapidly changing information generated by millions of users that can dramatically affect the reputation of a person or an organization. This raises the importance of automatic extraction of sentiments and opinions expressed in social media.

YouTube is a unique environment, just like Twitter, but probably even richer: multi-modal, with a social graph, and discussions between people sharing an interest. Hence, doing sentiment research in such an environment is highly relevant for the community. While the linguistic conventions used on Twitter and YouTube indeed show similarities (Baldwin et al., 2013), focusing on YouTube allows to exploit context information, possibly also multi-modal information, not available in isolated tweets, thus rendering it a valuable resource for the future research.

Nevertheless, there is almost no work showing effective OM on YouTube comments. To the best of our knowledge, the only exception is given by the classification system of YouTube comments proposed by Siersdorfer et al. (2010).

While previous state-of-the-art models for opinion classification have been successfully applied to traditional corpora (Pang and Lee, 2008), YouTube comments pose additional challenges: (i) polarity words can refer to either video or product while expressing contrasting sentiments; (ii) many comments are unrelated or contain spam; and (iii) learning supervised models requires training data for each different YouTube domain, e.g., tablets, automobiles, etc. For example, consider a typical comment on a YouTube review video about a Motorola Xoom tablet:

this guy really puts a negative spin on this, and I’m not sure why, this seems crazy fast, and I’m not entirely sure why his pinch to zoom is laggy all the other xoom reviews

The comment contains a product name xoom and some negative expressions, thus, a bag-of-words model would derive a negative polarity for this product. In contrast, the opinion towards the product is neutral as the negative sentiment is expressed towards the video. Similarly, the following comment:

iPad 2 is better. the superior apps just destroy the xoom.

contains two positive and one negative word, yet the sentiment towards the product is negative (the negative word destroy refers to Xoom). Clearly, the bag-of-words lacks the structural information linking the sentiment with the target product.

In this paper, we carry out a systematic study on OM targeting YouTube comments; its contribution is three-fold: firstly, to solve the problems outlined above, we define a classification schema, which separates spam and not related comments from the informative ones, which are, in turn, further categorized into video- or product-related comments

1252
corpus of blogs (Kessler et al., 2010), etc. The aforementioned corpora are, however, only partially suitable for developing models on social media, since the informal text poses additional challenges for Information Extraction and Natural Language Processing. Similar to Twitter, most YouTube comments are very short, the language is informal with numerous accidental and deliberate errors and grammatical inconsistencies, which makes previous corpora less suitable to train models for OM on YouTube. A recent study focuses on sentiment analysis for Twitter (Pak and Paroubek, 2010), however, their corpus was compiled automatically by searching for emotions expressing positive and negative sentiment only.

Siersdorfer et al. (2010) focus on exploiting user ratings (counts of ‘thumbs up/down’ as flagged by other users) of YouTube video comments to train classifiers to predict the community acceptance of new comments. Hence, their goal is different: predicting comment ratings, rather than predicting the sentiment expressed in a YouTube comment or its information content. Exploiting the information from user ratings is a feature that we have not exploited thus far, but we believe that it is a valuable feature to use in future work.

Most of the previous work on supervised sentiment analysis use feature vectors to encode documents. While a few successful attempts have been made to use more involved linguistic analysis for opinion mining, such as dependency trees with latent nodes (Täckström and McDonald, 2011) and syntactic parse trees with vectorized nodes (Socher et al., 2011), recently, a comprehensive study by Wang and Manning (2012) showed that a simple model using bigrams and SVMs performs on par with more complex models.

In contrast, we show that adding structural features from syntactic trees is particularly useful for the cross-domain setting. They help to build a system that is more robust across domains. Therefore, rather than trying to build a specialized system for every new target domain, as it has been done in most prior work on domain adaptation (Blitzer et al., 2007; Daumé, 2007), the domain adaptation problem boils down to finding a more robust system (Søgaard and Johannsen, 2012; Plank and Moschitti, 2013). This is in line with recent advances in parsing the web (Petrov and McDonald, 2012), where participants where asked to build a single system able to cope with different yet re-

2 Related work

Most prior work on more general OM has been carried out on more standardized forms of text, such as consumer reviews or newswire. The most commonly used datasets include: the MPQA corpus of news documents (Wilson et al., 2005), web customer review data (Hu and Liu, 2004), Amazon review data (Blitzer et al., 2007), the JDPA

(1) The corpus and the annotation guidelines are publicly available at http://projects.disi.unitn.it/iKernels/projects/sentube/
lated domains.

Our approach relies on robust syntactic structures to automatically generate patterns that adapt better. These representations have been inspired by the semantic models developed for Question Answering (Moschitti, 2008; Severyn and Moschitti, 2012; Severyn and Moschitti, 2013) and Semantic Textual Similarity (Severyn et al., 2013). Moreover, we introduce additional tags, e.g., video concepts, polarity and negation words, to achieve better generalization across different domains where the word distribution and vocabulary changes.

3 Representations and models

Our approach to OM on YouTube relies on the design of classifiers to predict comment type and opinion polarity. Such classifiers are traditionally based on bag-of-words and more advanced features. In the next sections, we define a baseline feature vector model and a novel structural model based on kernel methods.

3.1 Feature Set

We enrich the traditional bag-of-word representation with features from a sentiment lexicon and features quantifying the negation present in the comment. Our model (FVEC) encodes each document using the following feature groups:

- **word n-grams**: we compute unigrams and bigrams over lower-cased word lemmas where binary values are used to indicate the presence/absence of a given item.
- **lexicon**: a sentiment lexicon is a collection of words associated with a positive or negative sentiment. We use two manually constructed sentiment lexicons that are freely available: the MPQA Lexicon (Wilson et al., 2005) and the lexicon of Hu and Liu (2004). For each of the lexicons, we use the number of words found in the comment that have positive and negative sentiment as a feature.
- **negation**: the count of negation words, e.g., {don’t, never, not, etc.}, found in a comment. Our structural representation (defined next) enables a more involved treatment of negation.
- **video concept**: cosine similarity between a comment and the title/description of the video. Most of the videos come with a title and a short description, which can be used to encode the topicality of each comment by looking at their overlap.

3.2 Structural model

We go beyond traditional feature vectors by employing structural models (STRUCT), which encode each comment into a shallow syntactic tree. These trees are input to tree kernel functions for generating structural features. Our structures are specifically adapted to the noisy user-generated texts and encode important aspects of the comments, e.g., words from the sentiment lexicons, product concepts and negation words, which specifically targets the sentiment and comment type classification tasks.

In particular, our shallow tree structure is a two-level syntactic hierarchy built from word lemmas (leaves) and part-of-speech tags that are further grouped into chunks (Fig. 1). As full syntactic parsers such as constituency or dependency tree parsers would significantly degrade in performance on noisy texts, e.g., Twitter or YouTube comments, we opted for shallow structures, which rely on simpler and more robust components: a part-of-speech tagger and a chunker. Moreover, such taggers have been recently updated with models (Ritter et al., 2011; Gimpel et al., 2011) trained specifically to process noisy texts showing significant reductions in the error rate on user-generated texts, e.g., Twitter. Hence, we use the CMU Twitter pos-tagger (Gimpel et al., 2011; Owoputi et al., 2013) to obtain the part-of-speech tags. Our second component – chunker – is taken from (Ritter et al., 2011), which also comes with a model trained on Twitter data and shown to perform better on noisy data such as user comments.

To address the specifics of OM tasks on YouTube comments, we enrich syntactic trees with semantic tags to encode: (i) central concepts of the video, (ii) sentiment-bearing words expressing positive or negative sentiment and (iii) negation words. To automatically identify concept words of the video we use context words (tokens detected as nouns by the part-of-speech tagger) from the video title and video description and match them in the tree. For the matched words, we enrich labels of their parent nodes (part-of-speech and chunk) with the PRODUCT tag. Similarly, the nodes associated with words found in

---

2The list of negation words is adopted from http://sentiment.christopherpotts.net/lingstruc.html

3The chunker from (Ritter et al., 2011) relies on its own POS tagger, however, in our structural representations we favor the POS tags from the CMU Twitter tagger and take only the chunk tags from the chunker.
the sentiment lexicon are enriched with a polarity tag (either positive or negative), while negation words are labeled with the NEG tag. It should be noted that vector-based (FVEC) model relies only on feature counts whereas the proposed tree encodes powerful contextual syntactic features in terms of tree fragments. The latter are automatically generated and learned by SVMs with expressive tree kernels.

For example, the comment in Figure 1 shows two positive and one negative word from the sentiment lexicon. This would strongly bias the FVEC sentiment classifier to assign a positive label to the comment. In contrast, the STRUCT model relies on the fact that the negative word, destroy, refers to the PRODUCT (xoom) since they form a verbal phase (VP). In other words, the tree fragment:

\[ \text{[S [negative-VP [destroy]] [PRODUCT-NP [PRODUCT-N [xoom]]]]} \]

is a strong feature (induced by tree kernels) to help the classifier to discriminate such hard cases. Moreover, tree kernels generate all possible subtrees, thus producing generalized (back-off) features, e.g., \[ \text{[S [negative-VP [PRODUCT-NP]]]} \] or \[ \text{[S [negative-VP [PRODUCT-NP]]]} \].

3.3 Learning

We perform OM on YouTube using supervised methods, e.g., SVM. Our goal is to learn a model that automatically detects the sentiment and type of each comment. For this purpose, we build a multi-class classifier using the one-vs-all scheme. A binary classifier is trained for each of the classes and the predicted class is obtained by taking a class from the classifier with a maximum prediction score. Our back-end binary classifier is SVM-light-TK\(^4\), which encodes structural kernels in the SVM-light (Joachims, 2002) solver. We define a novel and efficient tree kernel function, namely, Shallow syntactic Tree Kernel (SHTK), which is as expressive as the Partial Tree Kernel (PTK) (Moschitti, 2006a) to handle feature engineering over the structural representations of the STRUCT model. A polynomial kernel of degree 3 is applied to feature vectors (FVEC).

Combining structural and vector models. A typical kernel machine, e.g., SVM, classifies a test input \( \mathbf{x} \) using the following prediction function:

\[ h(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i), \]

where \( \alpha_i \) are the model parameters estimated from the training data, \( y_i \) are target variables, \( \mathbf{x}_i \) are support vectors, and \( K(\cdot, \cdot) \) is a kernel function. The latter computes the similarity between two comments. The STRUCT model treats each comment as a tuple \( \mathbf{x} = (T, v) \) composed of a shallow syntactic tree \( T \) and a feature vector \( v \). Hence, for each pair of comments \( \mathbf{x}_1 \) and \( \mathbf{x}_2 \), we define the following comment similarity kernel:

\[ K(\mathbf{x}_1, \mathbf{x}_2) = K_{\text{TK}}(T_1, T_2) + K_v(v_1, v_2), \quad (1) \]

where \( K_{\text{TK}} \) computes SHTK (defined next), and \( K_v \) is a kernel over feature vectors, e.g., linear, polynomial, Gaussian, etc.

Shallow syntactic tree kernel. Following the convolution kernel framework, we define the new

\(^4\)http://disi.unitn.it/moschitti/Tree-Kernel.htm
SHTK function from Eq. 1 to compute the similarity between tree structures. It counts the number of common substructures between two trees \( T_1 \) and \( T_2 \) without explicitly considering the whole fragment space. The general equations for Convolution Tree Kernels is:

\[
TK(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1, n_2), \tag{2}
\]

where \( N_{T_1} \) and \( N_{T_2} \) are the sets of the \( T_1 \)’s and \( T_2 \)’s nodes, respectively and \( \Delta(n_1, n_2) \) is equal to the number of common fragments rooted in the \( n_1 \) and \( n_2 \) nodes, according to several possible definition of the atomic fragments.

To improve the speed computation of \( TK \), we consider pairs of nodes \((n_1, n_2)\) belonging to the same tree level. Thus, given \( H \), the height of the \textsc{struct} trees, where each level \( h \) contains nodes of the same type, i.e., chunk, POS, and lexical nodes, we define SHTK as the following:

\[
SHTK(T_1, T_2) = \sum_{h=1}^{H} \sum_{n_1 \in N^h_{T_1}} \sum_{n_2 \in N^h_{T_2}} \Delta(n_1, n_2), \tag{3}
\]

where \( N^h_{T_1} \) and \( N^h_{T_2} \) are sets of nodes at height \( h \).

The above equation can be applied with any \( \Delta \) function. To have a more general and expressive kernel, we use \( \Delta \) previously defined for PTK. More formally: if \( n_1 \) and \( n_2 \) are leaves then \( \Delta(n_1, n_2) = \mu \lambda (n_1, n_2) \); else \( \Delta(n_1, n_2) = \mu \left( \lambda^2 + \sum \lambda^{d(I_1)+d(I_2)} \prod_{j=1}^{\left| I_1 \right|} \Delta(c_{n_1}(I_{1j}), c_{n_2}(I_{2j})) \right) \),

where \( \lambda, \mu \in [0,1] \) are decay factors; the large sum is adopted from a definition of the subsequence kernel (Shawe-Taylor and Cristianini, 2004) to generate children subsets with gaps, which are then used in a recursive call to \( \Delta \). Here, \( c_{n_1}(i) \) is the \( i^{th} \) child of the node \( n_1 \); \( I_1 \) and \( I_2 \) are sequences of indexes that enumerate subsets of children with gaps, i.e., \( I = (i_1, i_2, ..., |I|) \), with \( 1 \leq i_1 < i_2 < ... < i_{|I|} \); and \( d(I_1) = \hat{I}_{11} - I_{11} + 1 \) and \( d(I_2) = \hat{I}_{21} - \hat{I}_{21} + 1 \), which penalizes subsequences with larger gaps.

It should be noted that: firstly, the use of a subsequence kernel makes it possible to generate child subsets of the two nodes, i.e., it allows for gaps, which makes matching of syntactic patterns less rigid. Secondly, the resulting SHTK is essentially a special case of PTK (Moschitti, 2006a), adapted to the shallow structural representation \textsc{struct} (see Sec. 3.2). When applied to \textsc{struct} trees, SHTK exactly computes the same feature space as PTK, but in faster time (on average). Indeed, SHTK required to be only applied to node pairs from the same level (see Eq. 3), where the node labels can match – chunk, POS or lexicals. This reduces the time for selecting the matching-node pairs carried out in PTK (Moschitti, 2006a; Moschitti, 2006b). The fragment space is obviously the same, as the node labels of different levels in \textsc{struct} are different and will not be matched by PTK either.

Finally, given its recursive definition in Eq. 3 and the use of subsequence (with gaps), SHTK can derive useful dependencies between its elements. For example, it will generate the following subtree fragments: \([\text{positive-NP} [\text{positive-ADN}]], [S [\text{negative-VP} [\text{negative-V [destroy]]} [\text{PRODUCT-NP}]]]\) and so on.

### 4 YouTube comments corpus

To build a corpus of YouTube comments, we focus on a particular set of videos (technical reviews and advertisings) featuring commercial products. In particular, we chose two product categories: automobiles (\textsc{auto}) and tablets (\textsc{tablets}). To collect the videos, we compiled a list of products and queried the YouTube gData API\(^6\) to retrieve the videos. We then manually excluded irrelevant videos. For each video, we extracted all available comments (limited to maximum 1k comments per video) and manually annotated each comment with its type and polarity. We distinguish between the following types:

- **product**: discuss the topic product in general or some features of the product;
- **video**: discuss the video or some of its details;
- **spam**: provide advertising and malicious links;
- **off-topic**: comments that have almost no content (“I’m a virus”) or content that is not related to the video (“Thank you!”).

Regarding the polarity, we distinguish between \{positive, negative, neutral\} sentiments with respect to the product and the video. If the comment contains several statements of different polarities, it is annotated as both positive and negative: “Love the video but waiting for iPad 4”. In total we have

\(^6\)https://developers.google.com/youtube/v3/
annotated 208 videos with around 35k comments (128 videos TABLETS and 80 for AUTO).

To evaluate the quality of the produced labels, we asked 5 annotators to label a sample set of one hundred comments and measured the agreement. The resulting annotator agreement α value (Krippendorf, 2004; Artstein and Poesio, 2008) scores are 60.6 (AUTO), 72.1 (TABLETS) for the sentiment task and 64.1 (AUTO), 79.3 (TABLETS) for the type classification task. For the rest of the comments, we assigned the entire annotation task to a singlecoder. Further details on the corpus can be found in Uryupina et al. (2014).

5 Experiments

This section reports: (i) experiments on individual subtasks of opinion and type classification; (ii) the full task of predicting type and sentiment; (iii) study on the adaptability of our system by learning on one domain and testing on the other; (iv) learning curves that provide an indication on the required amount and type of data and the scalability to other domains.

5.1 Task description

Sentiment classification. We treat each comment as expressing positive, negative or neutral sentiment. Hence, the task is a three-way classification.

Type classification. One of the challenging aspects of sentiment analysis of YouTube data is that the comments may express the sentiment not only towards the product shown in the video, but also the video itself, i.e., users may post positive comments to the video while being generally negative about the product and vice versa. Hence, it is of crucial importance to distinguish between these two types of comments. Additionally, many comments are irrelevant for both the product and the video (off-topic) or may even contain spam. Given that the main goal of sentiment analysis is to select sentiment-bearing comments and identify their polarity, distinguishing between off-topic and spam categories is not critical. Thus, we merge the spam and off-topic into a single uninformative category. Similar to the opinion classification task, comment type classification is a multi-class classification with three classes: video, product and uninform.

Full task. While the previously discussed sentiment and type identification tasks are useful to model and study in their own right, our end goal is: given a stream of comments, to jointly predict both the type and the sentiment of each comment. We cast this problem as a single multi-class classification task with seven classes: the Cartesian product {product, video} type labels and {positive, neutral, negative} sentiment labels plus the uninformative category (spam and off-topic). Considering a real-life application, it is important not only to detect the polarity of the comment, but to also identify if it is expressed towards the product or the video.7

5.2 Data

We split all the videos 50% between training set (TRAIN) and test set (TEST), where each video contains all its comments. This ensures that all comments from the same video appear either in TRAIN or in TEST. Since the number of comments per video varies, the resulting sizes of each set are different (we use the larger split for TRAIN). Table 1 shows the data distribution across the task-specific classes – sentiment and type classification. For the sentiment task we exclude off-topic and spam comments as well as comments with ambiguous sentiment, i.e., an-

<table>
<thead>
<tr>
<th>Task</th>
<th>class</th>
<th>AUTO</th>
<th>TABLETS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TRAIN</td>
<td>TEST</td>
<td>TRAIN</td>
</tr>
<tr>
<td>Sentiment</td>
<td>positive</td>
<td>2065 (35%)</td>
<td>807 (27%)</td>
</tr>
<tr>
<td></td>
<td>neutral</td>
<td>878 (16%)</td>
<td>760 (26%)</td>
</tr>
<tr>
<td></td>
<td>negative</td>
<td>8774 6960</td>
<td>5192</td>
</tr>
<tr>
<td>Total</td>
<td>11766</td>
<td>8927</td>
<td>11766</td>
</tr>
<tr>
<td>Type</td>
<td>product</td>
<td>2393 (19%)</td>
<td>1606 (17%)</td>
</tr>
<tr>
<td></td>
<td>video</td>
<td>1698 (18%)</td>
<td>1471 (21%)</td>
</tr>
<tr>
<td></td>
<td>off-topic</td>
<td>1002 (9%)</td>
<td>773 (9%)</td>
</tr>
<tr>
<td></td>
<td>spam</td>
<td>1648 (12%)</td>
<td>1278 (14%)</td>
</tr>
<tr>
<td>Total</td>
<td>3196</td>
<td>501</td>
<td>3196</td>
</tr>
</tbody>
</table>

Table 1: Summary of YouTube comments data used in the sentiment, type and full classification tasks. The comments come from two product categories: AUTO and TABLETS. Numbers in parenthesis show proportion w.r.t. to the total number of comments used in a task.

---

7We exclude comments annotated as both video and product. This enables the use of a simple flat multi-classifiers with seven categories for the full task, instead of a hierarchical multi-label classifiers (i.e., type classification first and then opinion polarity). The number of comments assigned to both product and video is relatively small (8% for TABLETS and 4% for AUTO).
notated as both positive and negative.

For the sentiment task about 50% of the
comments have neutral polarity, while the
negative class is much less frequent. Interest-
ingly, the ratios between polarities expressed
in comments from AUTO and TABLETS are very
similar across both TRAIN and TEST. Conversely,
for the type task, we observe that comments from
AUTO are uniformly distributed among the three
classes, while for the TABLETS the majority of
comments are product related. It is likely due
to the nature of the TABLETS videos, that are
more geek-oriented, where users are more prone
to share their opinions and enter involved discus-
sions about a product. Additionally, videos from
the AUTO category (both commercials and user
reviews) are more visually captivating and, be-
ing generally oriented towards a larger audience,
generate more video-related comments. Regarding
the full setting, where the goal is to have
a joint prediction of the comment sentiment and
type, we observe that video-negative and
video-positive are the most scarce classes,
which makes them the most difficult to predict.

5.3 Results

We start off by presenting the results for the tradi-
tional in-domain setting, where both TRAIN and
TEST come from the same domain, e.g., AUTO or
TABLETS. Next, we show the learning curves to
analyze the behavior of FVEC and STRUCT mod-
els according to the training size. Finally, we per-
form a set of cross-domain experiments that de-
scribe the enhanced adaptability of the patterns
generated by the STRUCT model.

5.3.1 In-domain experiments

We compare FVEC and STRUCT models on three
tasks described in Sec. 5.1: sentiment, type and
full. Table 2 reports the per-class performance
and the overall accuracy of the multi-class clas-
sifier. Firstly, we note that the performance on
TABLETS is much higher than on AUTO across
all tasks. This can be explained by the follow-
ing: (i) TABLETS contains more training data and
(ii) videos from AUTO and TABLETS categories
draw different types of audiences – well-informed
users and geeks expressing better-motivated opin-
ions about a product for the former vs. more gen-
eral audience for the latter. This results in the
different quality of comments with the AUTO be-
ing more challenging to analyze. Secondly, we
observe that the STRUCT model provides 1-3%
of absolute improvement in accuracy over FVEC
for every task. For individual categories the F1
scores are also improved by the STRUCT model
(except for the negative classes for AUTO, where
we see a small drop). We conjecture that sentiment
prediction for AUTO category is largely driven
by one-shot phrases and statements where it is
hard to improve upon the bag-of-words and senti-
ment lexicon features. In contrast, comments from
TABLETS category tend to be more elaborated
and well-argumented, thus, benefiting from the ex-
pressiveness of the structural representations.

Considering per-class performance, correctly
predicting negative sentiment is most difficult
for both AUTO and TABLETS, which is proba-
bly caused by the smaller proportion of the neg-
ative comments in the training set. For the type
task, video-related class is substantially more dif-
ficult than product-related for both categories. For
the full task, the class video-negative ac-
counts for the largest error. This is confirmed by
the results from the previous sentiment and type
tasks, where we saw that handling negative sen-
timent and detecting video-related comments are
most difficult.

5.3.2 Learning curves

The learning curves depict the behavior of FVEC
and STRUCT models as we increase the size of
the training set. Intuitively, the STRUCT model
relies on more general syntactic patterns and may
overcome the sparseness problems incurred by the
FVEC model when little training data is available.

Nevertheless, as we see in Figure 2, the learning
curves for sentiment and type classification tasks
across both product categories do not confirm this
intuition. The STRUCT model consistently outper-
forms the FVEC across all training sizes, but the
gap in the performance does not increase when we
move to smaller training sets. As we will see next,
this picture changes when we perform the cross-
domain study.

5.3.3 Cross-domain experiments

To understand the performance of our classifiers
on other YouTube domains, we perform a set of
cross-domain experiments by training on the data
from one product category and testing on the other.

Table 3 reports the accuracy for three tasks
when we use all comments (TRAIN + TEST) from
AUTO to predict on the TEST from TABLETS.
Table 2: In-domain experiments on AUTO and TABLETS using two models: FVEC and STRUCT. The results are reported for sentiment, type and full classification tasks. The metrics used are precision (P), recall (R) and F1 for each individual class and the general accuracy of the multi-class classifier (Acc).

<table>
<thead>
<tr>
<th>Task</th>
<th>class</th>
<th>AUTO</th>
<th>TABLETS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FVEC</td>
<td>STRUCT</td>
<td>FVEC</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>Sent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>positive</td>
<td>49.1</td>
<td>72.1</td>
<td>58.4</td>
</tr>
<tr>
<td>neutral</td>
<td>68.2</td>
<td>55.0</td>
<td>61.4</td>
</tr>
<tr>
<td>negative</td>
<td>42.0</td>
<td>36.9</td>
<td>39.6</td>
</tr>
<tr>
<td>Acc</td>
<td>54.7</td>
<td>55.7</td>
<td>68.6</td>
</tr>
<tr>
<td>Type</td>
<td>product</td>
<td>66.8</td>
<td>73.3</td>
</tr>
<tr>
<td></td>
<td>video</td>
<td>45.0</td>
<td>52.8</td>
</tr>
<tr>
<td></td>
<td>uninform</td>
<td>59.3</td>
<td>48.2</td>
</tr>
<tr>
<td>Acc</td>
<td>57.4</td>
<td>59.4</td>
<td>77.2</td>
</tr>
<tr>
<td>Full</td>
<td>product-pos</td>
<td>34.0</td>
<td>49.6</td>
</tr>
<tr>
<td></td>
<td>product-neu</td>
<td>43.4</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td>video-pos</td>
<td>23.2</td>
<td>47.1</td>
</tr>
<tr>
<td></td>
<td>video-neu</td>
<td>26.1</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>video-neg</td>
<td>21.9</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>uninform</td>
<td>56.5</td>
<td>52.4</td>
</tr>
<tr>
<td>Acc</td>
<td>40.0</td>
<td>41.5</td>
<td>57.6</td>
</tr>
</tbody>
</table>

Figure 2: In-domain learning curves. ALL refers to the entire TRAIN set for a given product category, i.e., AUTO and TABLETS (see Table 1) and in the opposite direction (TABLETS→AUTO). When using AUTO as a source domain, STRUCT model provides additional 1-3% of absolute improvement, except for the sentiment task.

Similar to the in-domain experiments, we studied the effect of the source domain size on the target test performance. This is useful to assess the adaptability of features exploited by the FVEC and STRUCT models with the change in the number of labeled examples available for training. Additionally, we considered a setting including a small amount of training data from the target data (i.e., supervised domain adaptation).

For this purpose, we drew the learning curves of the FVEC and STRUCT models applied to the sentiment and type tasks (Figure 3): AUTO is used as the source domain to train models, which are tested on TABLETS.8 The plot shows that when

Table 3: Cross-domain experiment. Accuracy using FVEC and STRUCT models when trained/tested in both directions, i.e., AUTO→TABLETS and TABLETS→AUTO. † denotes results statistically significant at 95% level (via pairwise t-test).

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Task</th>
<th>FVEC</th>
<th>STRUCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUTO</td>
<td>TABLETS</td>
<td>Sent</td>
<td>66.1</td>
<td>66.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type</td>
<td>59.9</td>
<td>64.1†</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Full</td>
<td>35.6</td>
<td>38.3†</td>
</tr>
<tr>
<td>TABLETS</td>
<td>AUTO</td>
<td>Sent</td>
<td>60.4</td>
<td>61.9†</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type</td>
<td>54.2</td>
<td>55.6†</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Full</td>
<td>43.4</td>
<td>44.7†</td>
</tr>
</tbody>
</table>

8The results for the other direction (TABLETS→AUTO) show similar behavior.
little training data is available, the features generated by the STRUCT model exhibit better adaptability (up to 10% of improvement over FVEC). The bag-of-words model seems to be affected by the data sparsity problem which becomes a crucial issue when only a small training set is available. This difference becomes smaller as we add data from the same domain. This is an important advantage of our structural approach, since we cannot realistically expect to obtain manual annotations for 10k+ comments for each (of many thousands) product domains present on YouTube.

5.4 Discussion

Our STRUCT model is more accurate since it is able to induce structural patterns of sentiment. Consider the following comment: *optimus pad is better. this xoom is just to bulky but optimus pad offers better functionality*. The FVEC bag-of-words model misclassifies it to be positive, since it contains two positive expressions (*better, better functionality*) that outweigh a single negative expression (*bulky*). The structural model, in contrast, is able to identify the product of interest (*xoom*) and associate it with the negative expression through a structural feature and thus correctly classify the comment as negative.

Some issues remain problematic even for the structural model. The largest group of errors are implicit sentiments. Thus, some comments do not contain any explicit positive or negative opinions, but provide detailed and well-argued criticism, for example, *this phone is heavy*. Such comments might also include irony. To account for these cases, a deep understanding of the product domain is necessary.

6 Conclusions and Future Work

We carried out a systematic study on OM from YouTube comments by training a set of supervised multi-class classifiers distinguishing between video and product related opinions. We use standard feature vectors augmented by shallow syntactic trees enriched with additional conceptual information.

This paper makes several contributions: (i) it shows that effective OM can be carried out with supervised models trained on high quality annotations; (ii) it introduces a novel annotated corpus of YouTube comments, which we make available for the research community; (iii) it defines novel structural models and kernels, which can improve on feature vectors, e.g., up to 30% of relative improvement in type classification, when little data is available, and demonstrates that the structural model scales well to other domains.

In the future, we plan to work on a joint model to classify all the comments of a given video, s.t. it is possible to exploit latent dependencies between entities and the sentiments of the comment thread. Additionally, we plan to experiment with hierarchical multi-label classifiers for the full task (in place of a flat multi-class learner).

Acknowledgments

The authors are supported by a Google Faculty Award 2011, the Google Europe Fellowship Award 2013 and the European Community’s Seventh Framework Programme (FP7/2007-2013) under the grant #288024: LiMoSINE.
References


Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how different social media sources? In *IJCNLP*.


Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*.


Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

Barbara Plank and Alessandro Moschitti. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *ACL*.


