YOUCAT : WEAKLY SUPERVISED YOUTUBE VIDEO CATEGORIZATION SYSTEM FROM META DATA & USER COMMENTS USING WORDNET & WIKIPEDIA

Subhabrata Mukherjee[1,2], Pushpak Bhattacharyya[2]

IBM Research Lab, India[1]

Dept. of Computer Science and Engineering, IIT Bombay[2]

24th International Conference on Computational Linguistics

COLING 2012,

IIT Bombay, Mumbai, Dec 8 - Dec 15, 2012

Motivation

- In recent times, there has been an explosion in the number of online videos
- Efficient query-based information retrieval has become very important for the multimedia content
- For this, the genre or category identification of the video is essential
- Genre identification has been traditionally posed as a supervised classification task

Supervised Classification Issues

- A *serious challenge* for supervised classification in video categorization is the collection of manually labeled data (Filippova *et. al*, 2010; Wu *et. al*, 2010; Zanetti *et. al*, 2008)

- Consider a video with a descriptor "*It's the NBA's All-Mask Team!*"

- There must be a video in the training set with *NBA* in the video descriptor **labeled** with *Sport*, to identify its genre

- With increasing number of genres and incorporation of new genre-related concepts, data requirement rises

Supervised Classification Issues Contd...

- As new genres are introduced, labeled training data is required for the new genre
- *Very short text* is provided by the user, for title and video description, which provide little information (Wu *et. al,* 2010)
- Thus, video descriptors need to be expanded for better classification *or else feature space will be sparse*

Video Categorization Issues

- Title is very short
- Video descriptor is often very short or missing
- User comments are very noisy
  - Too many *slangs, abuses, off-topic conversations etc.*
- Requirement of a larger labeled training dataset

Novelty

- All the surveyed works are supervised with a lot of training data requirement
  - YouCat does not have any labeled data requirement
  - New genres can be easily introduced
- Use of WordNet and Wikipedia together has not been probed much for genre identification tasks

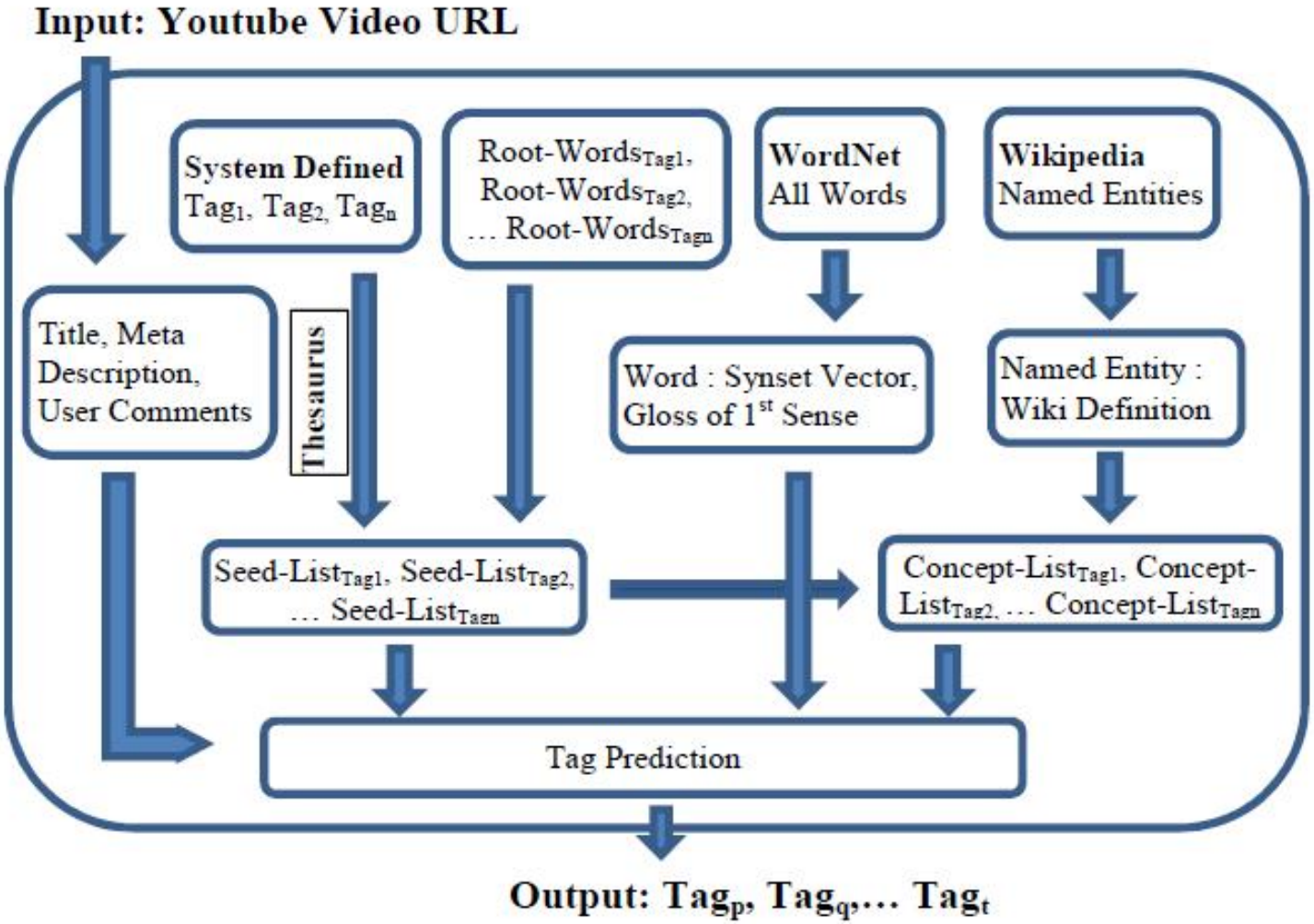
Questions

- We try to address the following questions in this work:
- Can a system without any labeled data requirement have comparable F-Score to a supervised system in genre identification tasks?
- Can incorporation of user comments, which are typically noisy, improve classification performance?
- Can the incorporation of lexical knowledge though WordNet and world knowledge through Wikipedia help in genre identification?
- Is it possible for the system to achieve all these requirements with a minimum time complexity, which is essential for real-time inter-active systems?

YouCat : Features

- Weakly supervised, requiring no labeled training data
  - Weak supervision arises out of
    - The usage of WordNet which is manually annotated
    - The specification of a set of 2-3 root words for each genre
- Introduction of new genres does not require training data
- Harvests features from *Video Title, Meta-Description and User Comments*
- Uses WordNet for lexical knowledge and Wikipedia for named entity information
- The genre identification algorithm has a time complexity of O(|W|), where |W| is the number of words in the video descriptor
- Does not use user-provided genre information and co-watch data

YouCat System Architecture



Genre Definition

- Genres are defined in the system beforehand with:
  - **Genre Name** – Example : *Comedy*
  - A set of **Root Words** (~ 2 – 3 words) for each genre which captures the characteristics of that genre – Example : *Laugh, Funny*

Comedy	**comedy, funny, laugh**
Horror	**horror, fear, scary**
Romance	**romance, romantic**
Sport	**sport, sports**
Technology	**tech, technology, science**

Table 1. Root Words for Each Genre

Step 1: Seed List Creation for Each Genre

- A **Seed List** of words is automatically created for each genre which captures all the key characteristics of that category.
  - Example - "*love", "hug", "cuddle" etc.* are the characteristics of the *Romance* genre

- *Root words of the genre* are taken and all their synonyms are retrieved from a thesaurus

- A thesaurus is used for this purpose which gives every day words and slangs
  - www.urbandictionary.com/thesaurus.php

Automatically Created Seed Word List from Thesaurus

Comedy (25)	funny, humor, hilarious, joke, comedy, roflmao, laugh, lol, rofl, roflmao, joke, giggle, haha, prank
Horror (37)	horror, curse, ghost, scary, zombie, terror, fear shock, evil, devil, creepy, monster, hell, blood, dead, demon
Romance (21)	love, romantic, dating, kiss, relationships, heart, hug, sex, cuddle, snug, smooch, crush, making out
Sports (35)	football, game, soccer, basketball, cheerleading, sports, baseball, FIFA, swimming, chess, cricket, shot
Tech (42)	internet, computers, apple, iPhone, phone, pc, laptop, mac, iPad, online, google, mac, laptop, XBOX, Yahoo

Table 2. Seed Words for Each Genre

Step 2.1: Concept Hashing (WordNet)

- Each word in WordNet is mapped to a genre using its **synset** and **gloss of first sense**

- Example - *dunk* has the synsets - {*dunk, dunk shot, stuff shot* ; *dunk*, *dip, souse, plunge, douse* ; *dunk* ; *dunk, dip*}.
- Gloss of the synset {*dunk, dunk shot, stuff shot*} is {*a basketball shot in which the basketball is propelled downward into the basket*}

- The words in the synset of its most appropriate sense, from the context, should have been taken

  It requires WSD

Step 2.2: Concept Hashing (Wikipedia)

- Named Entities are not present in the WordNet
- Wikipedia is necessary for *named entity* expansion
- All the *named entities* in Wikipedia with the *top 2 line definition* in their corresponding Wiki articles are retrieved.
- Example: *NBA* is retrieved from the Wikipedia article as {*The National Basketball Association (NBA) is the pre-eminent men's professional basketball league in North America. It consists of thirty franchised member clubs, of which twenty-nine are located in the United States and one in Canada.*}.
- A rough heuristic based on *Capitalization* is used to detect named entities (unigrams, bigrams, trigrams etc.)

Step 3: Concept List Creation

- Let **w** be any given word and its *expanded form* given by *WordNet* or *Wikipedia* be denoted by $\mathbf{w}'$.

- Let $\mathbf{w}'_\mathbf{j}$ be the j^{th} word in the expanded word vector. Let $\mathbf{seed_k}$ and $\mathbf{root_k}$ be the seed list and root words list, respectively, corresponding to the k^{th} genre.

- The genre of **w** is given by

$$genre(w) = argmax_k \sum_j \mathbf{1}_{w'_j \in seed_k, w'_j \in root_k}$$

- Here, $\mathbf{1}$ is an indicator function which returns 1 if a particular word is present in the seed list or root words list corresponding to a specific genre and 0 otherwise.

Step 3: Concept List Creation Contd…

- *Example : dunk* (from WordNet) and *NBA* (from Wikipedia) will be classified to the *Sports* genre as they have the maximum matches (“*shot”, “basketball”*) from the seed list corresponding to the *Sports* genre in their expanded concept vector.

- A *concept list* is created for each genre containing *associated* words in the WordNet and named entities in the Wikipedia

Video Descriptor Extraction

- Given a video url the *video title*, *meta description* of the video and the *user comments* on the video from Youtube are retrieved
- A *stopwords* list is used to remove words like *is, are, been etc*.
- A lemmatizer is used to reduce each word to its base form
  - Thus "*play*", "*played*", "*plays*", "*playing*" are reduced to its lemma "*play*"
- A *Feature Vector* is formed with all the extracted, lemmatized words in the video descriptor

Feature Vector Classification

- Let the video descriptor **f** consist of **n** words, in which the j^{th} word is denoted by $\mathbf{word_j}$.
- The *root word list, seed list* and the *concept list* for the k^{th} genre are denoted by $\mathbf{root_k}$, $\mathbf{seed_k}$ and $\mathbf{concept_k}$ respectively.

- The score of f belonging to a particular $genre_k$ is given by,

$$score(f \in genre_k; w_1, w_2, w_3) =$$
$$w_1 \times \sum_j \mathbf{1}_{word_j \in root_k} + w_2 \times \sum_j \mathbf{1}_{word_j \in seed_k} + w_3 \times \sum_j \mathbf{1}_{word_j \in concept_k}$$
$$where \quad w_3 < w_2 < w_1$$

- Here, $\mathbf{1}$ is an indicator function that returns 1 if a word is present in the root words list, seed list or concept list corresponding to $genre_k$ and 0 otherwise.

- Weights w_1, w_2 and w_3 are assigned to words present in the root words list, seed list and the concept list respectively.

Feature Vector Classification Contd…

- Weight assigned to any root word is maximum as it is specified, as part of the genre description, *manually*
- Lesser weightage is given to words in seed list, as they are automatically extracted using a thesaurus
- Weight assigned to concept list is the least to reduce the effect of *topic drift* during concept expansion
- The topic drift occurs due to enlarged context window, during concept expansion, which may result in a match from seed list of some other genre

Feature Vector Classification Contd…

- The score of a video belonging to a particular genre is,

$$score(video \in genre_k; p_1, p_2, p_3) = p_1 \times score(f^{Title} \in genre_k) + p_2 \times score(f^{Meta\ Data} \in genre_k) + p_3 \times score(f^{Comments} \in genre_k)$$

- Here p_1, p_2, p_3 denote the weight of the feature belonging to the *title, meta data* and *user comments* respectively where $p_1 > p_2 > p_3$
  - More importance is given to the title, then to the meta data and finally to the user comments

Feature Vector Classification Contd…

- The genre to which the video belongs is given by,

$$video_{genre} = argmax_k \; score(video \in genre_k)$$

- This assigns the highest scoring genre as the desired category for the video
- Most of the popular videos in Youtube can be attributed to more than one *genre*
- To allow multiple tags to be assigned to a video, a thesholding is done and the prediction is modified as:

$$video_{genre} = k, if \; score(video \in genre_k) \geq \theta$$

$$where \; \theta = \frac{1}{k} \sum_k score(video \in genre_k)$$

- If the genre scores for the 5 categories are something like {*400, 200, 100, 50, 10*} with *avg=152*, then the first 2 genres are chosen

Algorithm for Genre Identification

Pre-processing:

1.Define Genres and Root Words List for each genre
2.Create a Seed list for each genre by breadth-first-search in a Thesaurus, using root words in the genre or the genre name
3.Create a Concept List for each genre using all the words in WordNet (not present in Seed Lists) and Named Entities in Wikipedia using Equation 1

Input: Youtube Video Url

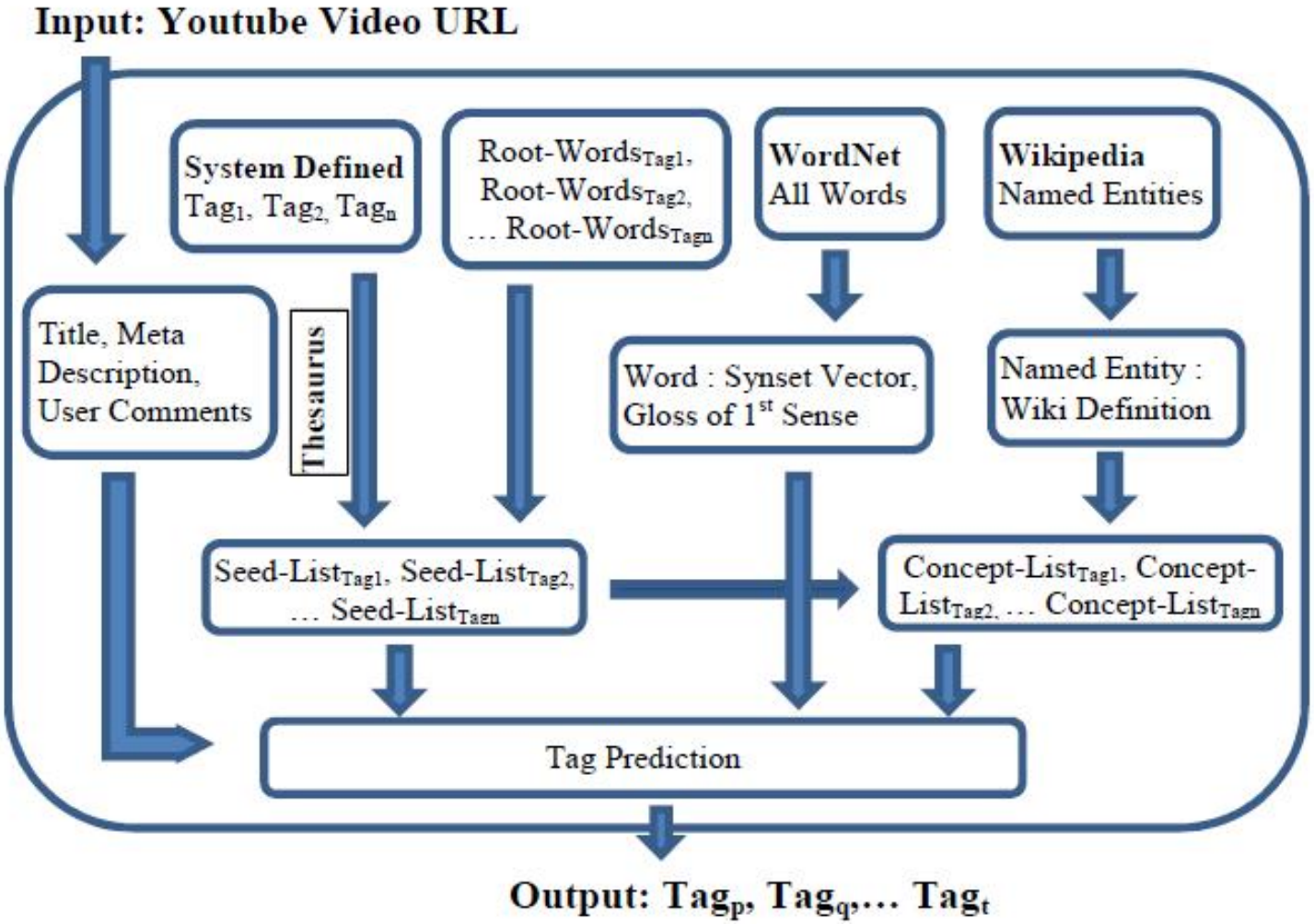
1.Extract Title, Meta Description of the video and User Comments from Youtube to form the video descriptor
2.Lemmatize all the words in the descriptor removing stop word.
3.Use Equations 2-4 for genre identification of the given video

Output: Genre Tags

Algorithm 1. Genre Identification of a Youtube Video

Time Complexity O(|W|), |W| is the number of words in the video descriptor

YouCat System Architecture



Parameter Setting: Unsupervised System

- Upweighting of document zones is common in ATS and IR
- Common strategy is to use extra weight for words appearing in certain portions of the text like the title
- As a rule-of-thumb the weights can be set as simple integral multiples, preferably prime, to reduce the possibility of ties (Manning, 2008)
- We upweight certain portions of the text like the *title, meta data*, *user comments* and assign different weight to words belonging to different lists according to importance.
- There are 6 parameters for the model we used: $w_1, w_2, w_3, p_1, p_2, p_3$. In the absence of any label information, we took the first set of integers, satisfying all the constraints in the *Equations,* and assigned them to the 6 parameters: $w_1 = 3, w_2 = 2, w_3 = 1, p_1 = 3, p_2 = 2, p_3 = 1.$

Parameter Setting: Partially Supervised System

- *The Equations can be written as:*

$$score\left(f_k^{position} \in genre_k; w_1, w_2, w_3\right) = w_1 \times X_{1,k}^{position} + w_2 \times X_{2,k}^{position} + w_3 \times X_{3,k}^{position}$$

$$score(video_k \in genre_k; p_1, p_2, p_3) = Y_k = \sum_{position} p_{position} \sum_j w_j \times X_{j,k}^{position}$$

$$= \sum_i \sum_j w'_{ij} X_j^i \quad (where\ w'_{ij} = p_i \times w_j)$$

$$Or, Y_k = \boldsymbol{W}.\boldsymbol{X}_k\ (where\ \boldsymbol{W} = \left[w'_{1,1}\ w'_{1,2} \ldots w'_{3,3}\right]^T_{9\times 1}, \qquad \boldsymbol{X}_k = \left[X^1_{1,k}\ X^1_{2,k} \ldots X^3_{3,k}\right]_{1\times 9}$$

$$Or, \boldsymbol{Y} = \boldsymbol{W}^{\boldsymbol{T}}.\boldsymbol{X}$$

- The solution for $\boldsymbol{W}$ is given by $\boldsymbol{W} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

A regularizer can be added to protect against over-fitting and the solution can be modified as: $\boldsymbol{W} = (\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ *where* δ *is a parameter and* $\boldsymbol{I}$ *is the identity matrix.*

Data Collection

- The following 5 genres are used for evaluation: *Comedy, Horror, Sports, Romance and Technology*

- 12,837 videos are crawled from the Youtube following a similar approach like Song *et al*. (2009), Cui *et al*. (2010) and Wu *et al.* (2012)

- Youtube has 15 pre-defined categories like *Romance, Music, Sports, People, Comedy* etc.
- These are categorized in Youtube based on the user-provided tags while uploading the video

- Videos are crawled from these categories and tags are verified

Data Collection Contd…

- Only the 1st page of user comments is taken
- Comments with length less than 150 characters in length are retained
- Only those user comments are taken whose support is greater than some threshold
- User comments are normalized by removing all the punctuations and reducing words like "*loveeee*" to "*love*"
- The number of user comments varied from 0 to 800 for different videos

Data Collection Contd…

Comedy	Horror	Sports	Romance	Tech	Total
2682	2802	2577	2477	2299	12837

Table 3. Number of Videos in Each Genre

Comedy	Horror	Sports	Romance	Tech
226	186	118	233	245

Table 4. Average User Comments for Each Genre

Baseline System

- **Multi-Class Support Vector Machines** Classifier with various features, like combination of unigrams and bigrams, incorporating part-of-speech (POS) information, removing stop words, using lemmatization *etc.*, is taken as the baseline

SVM Features	F_1-Score(%)
All Unigrams	82.5116
Unigrams+Without stop words	83.5131
Unigrams+ Without stop words +Lemmatization	83.8131
Unigrams+Without stop words +Lemmatization+ POS Tags	83.8213
Top Unigrams+Without stop words +Lemmatization+POS Tags	84.0524
All Bigrams	74.2681
Unigrams+Bigrams+Without Stop Words+Lemmatization	**84.3606**

Table 5: Multi-Class SVM Baseline with Different Features

Discussions: Multi-class SVM Baseline

- Ignoring stop words and lemmatization improves the accuracy of SVM
  - Related unigram features like *laugh, laughed, laughing* etc. are considered as a single entry *laugh,* which reduces sparsity of the feature space
- POS info increases accuracy, due to *crude* word sense disambiguation
- Consider the word *haunt* which has a noun sense - **a frequently visited place** and a verb sense - **follow stealthily or recur constantly and spontaneously to; her ex-boyfriend stalked her; the ghost of her mother haunted her**.
- The second sense is related to the *Horror* genre which can only be differentiated using POS tags.
- Top unigrams help in pruning the feature space and removing noise which helps in accuracy improvement
- Using bigrams along with unigrams gives the highest accuracy

YouCat Evaluation

- Experiments are performed on the videos *with* and *without user comments*, to find out whether user comments really assist in genre identification
- Experiments are also performed *with* and *without concept expansion*, to find out if WordNet and Wikipedia help in video categorization

- Performance is evaluated in terms of:

$$precision = \frac{number\ of\ videos\ correctly\ tagged}{number\ of\ videos\ tagged} \times 100$$

$$recall = \frac{number\ of\ video\ correctly\ tagged}{number\ of\ videos\ present\ in\ the\ genre} \times 100$$

$$f_1\ score = \frac{2 * precision * recall}{precision + recall}$$

Single Genre Prediction : Effect of User Comments

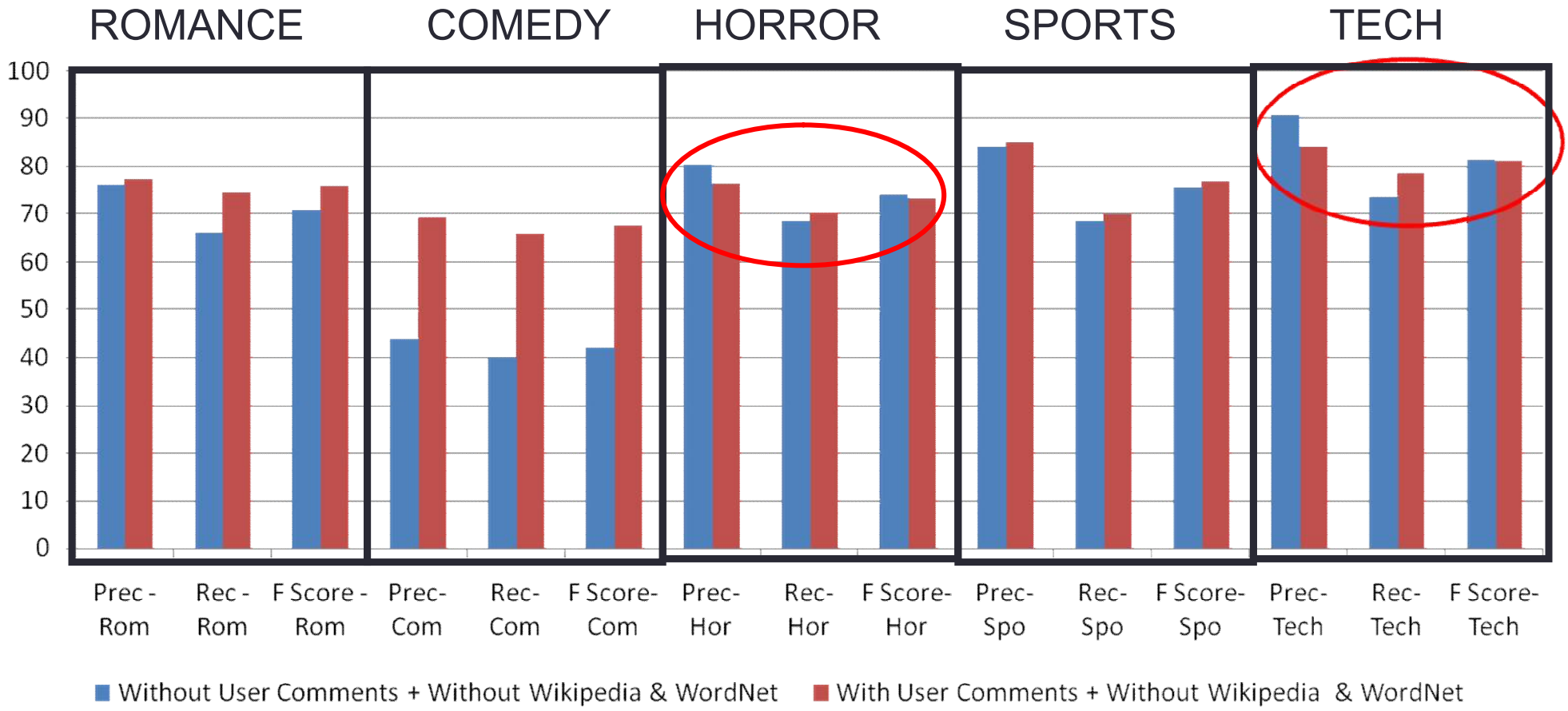


Table 6: Single Genre Identification with and without User Comments, without using Wikipedia & WordNet

Discussions : Effect of User Comments

- User comments introduce noise through off-topic conversations, spams, abuses *etc.*
- Slangs, abbreviations and pragmatics in user posts make analysis difficult
- Greater context provided by the user comments provide more clues about the genre
- User information mostly helps in identifying funny videos and romantic videos to some extent
- Horror videos undergo mild performance degradation

Single Genre Prediction : Effect of Concept Expansion

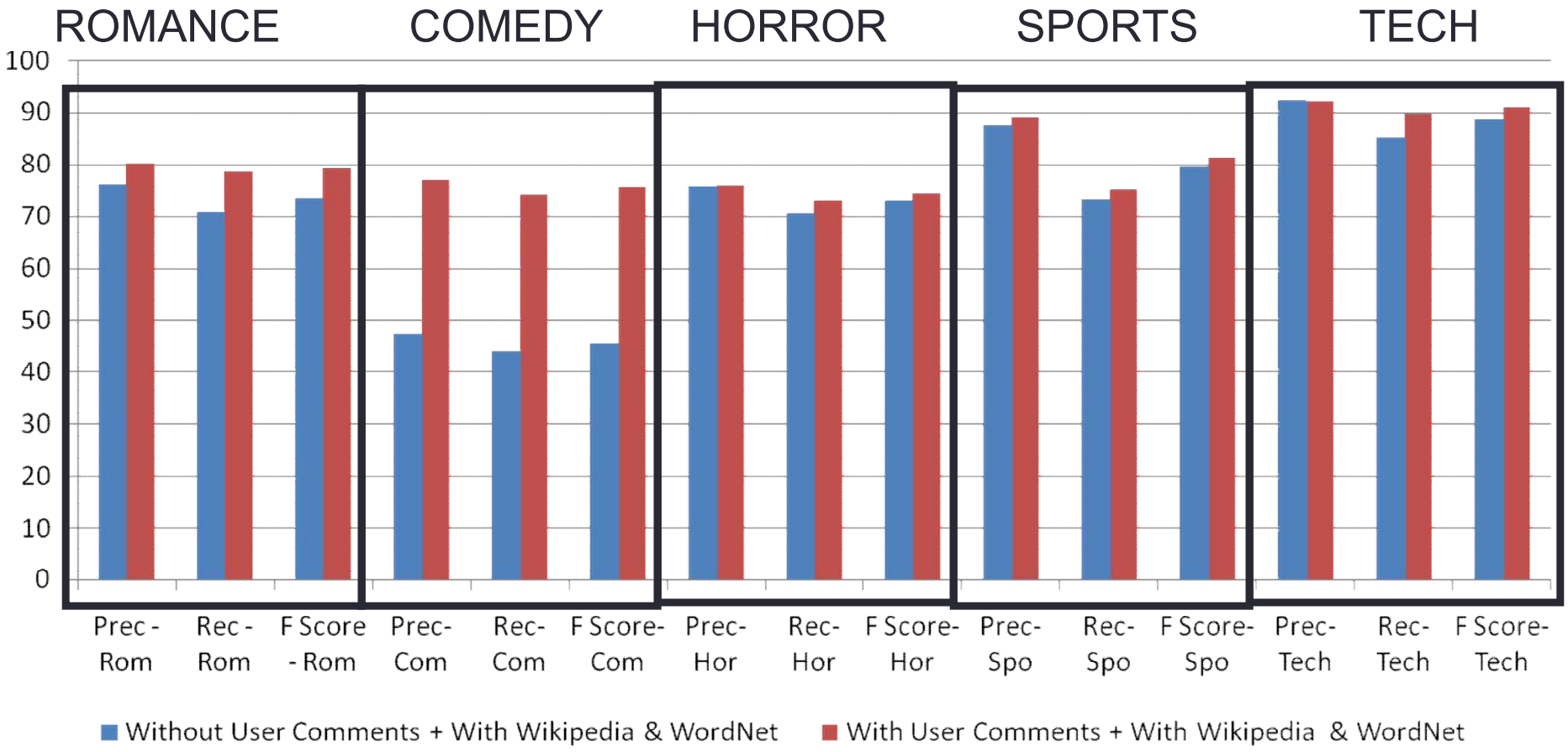


Table 7: Single Genre Identification with and without User Comments, using Wikipedia &WordNet

Discussions : Effect of Concept Expansion

- In single genre prediction f_1 score improvement of **3%** (when user comments are not used) and **6%** (when user comments are used) show that concept expansion is indeed helpful
- External knowledge sources help in easy identification of new **technological** concepts.
- Horror videos, again, undergo mild performance degradation
- Performance improvement in Comedy using Wikipedia can be attributed to the identification of the concepts like *Rotfl, Lolz, Lmfao etc*.

Multiple Genre Prediction : Effect of User Comments using Concept Expansion

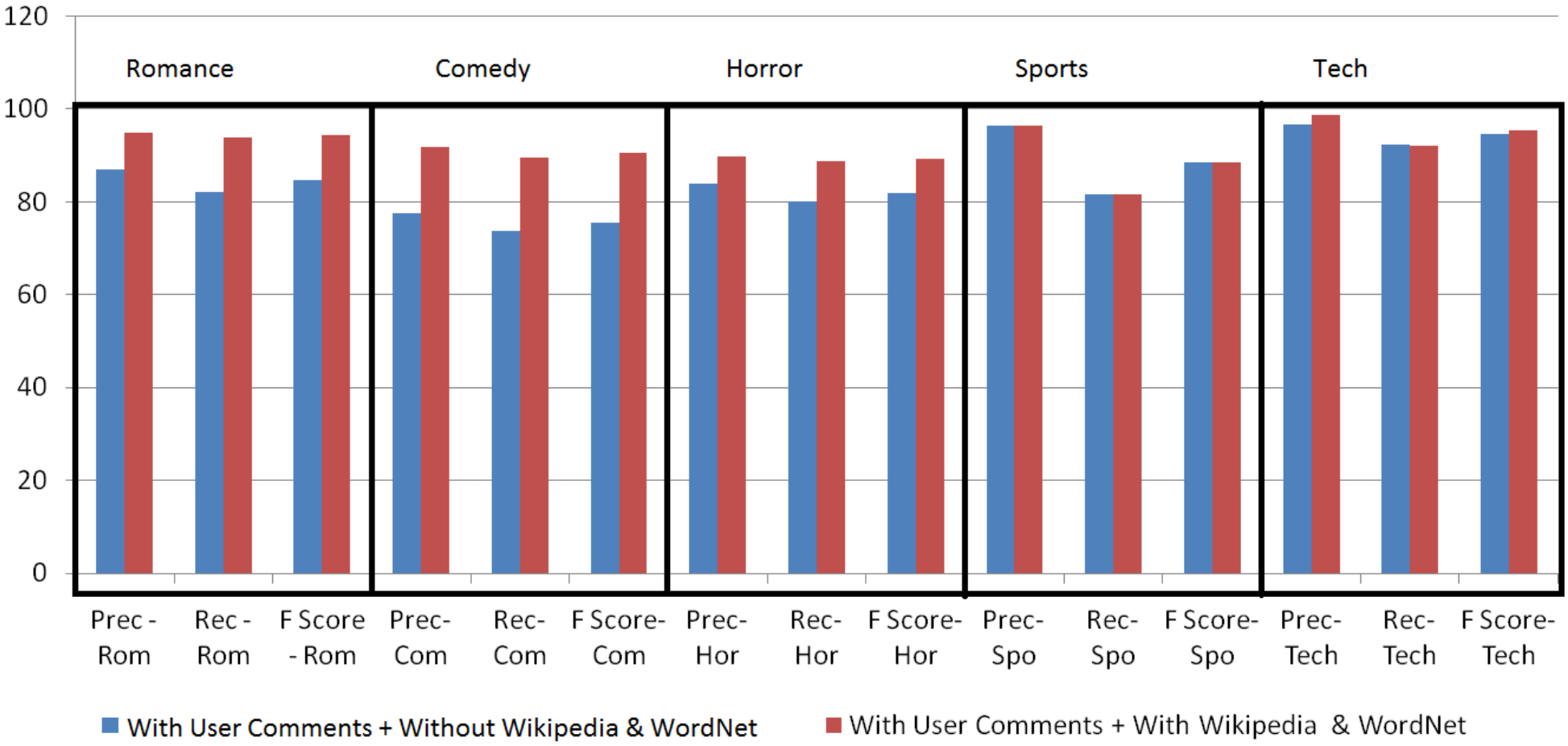
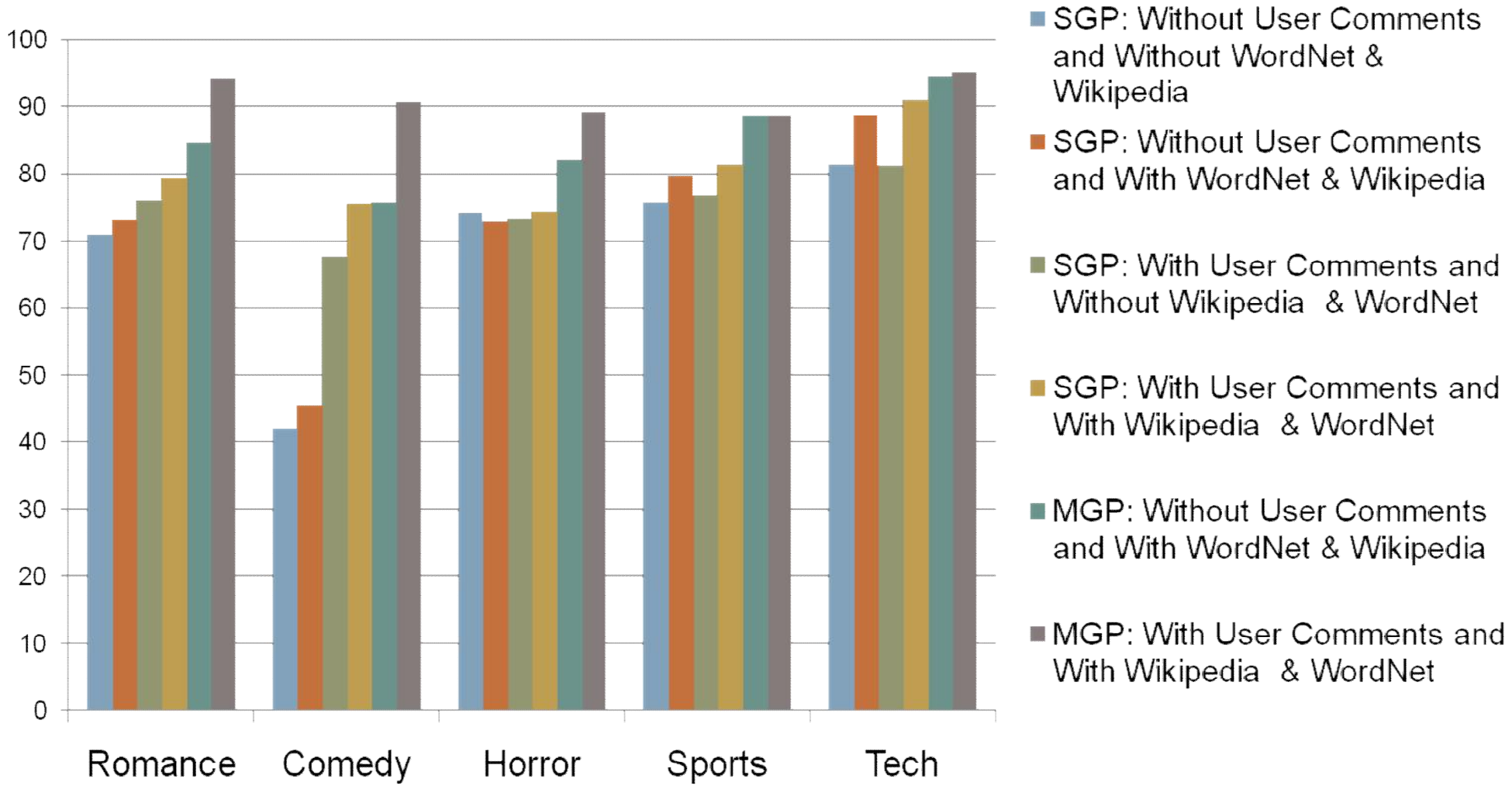


Table 8: Multiple Genre Identification with and without User Comments, using Wikipedia &WordNet

Genre-wise Results



Effect of Concept Expansion and User Comments on Single and Multiple Genre Prediction

Genre	Without User Comments + Without Wikipedia & WordNet			With User Comments + Without Wikipedia & WordNet		
	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Romance	76.26	66.27	70.91	77.36	74.60	**75.95**
Comedy	43.96	40.00	41.89	69.23	66.00	**67.58**
Horror	80.47	68.67	74.10	76.45	70.33	**73.26**
Sports	84.21	68.71	75.67	85.07	69.94	**76.77**
Tech	90.83	73.50	81.25	84.09	78.45	**81.17**

Genre	Without User Comments + With Wikipedia & WordNet			With User Comments + With Wikipedia & WordNet		
	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Romance	76.06	70.63	73.24	80.16	78.57	**79.36**
Comedy	47.31	44.00	45.6	77.08	74.00	**75.51**
Horror	75.63	70.33	72.88	75.78	73.00	**74.36**
Sports	87.5	73.01	79.60	89.05	74.85	**81.33**
Tech	92.34	85.16	88.60	92.03	89.75	**90.88**

Genre	Without User Comments + With Wikipedia & WordNet			With User Comments + With Wikipedia & WordNet		
	Precision	Recall	F_1-Score	Precision	Recall	F_1-Score
Romance	86.97	82.14	84.49	94.78	93.65	**94.21**
Comedy	77.5	73.67	75.54	91.78	89.33	**90.54**
Horror	83.92	80.00	81.91	89.56	88.67	**89.11**
Sports	96.38	81.6	88.38	96.38	81.6	**88.38**
Tech	96.63	92.34	94.44	98.56	92.03	**95.18**

Average Predicted Tags/Video in Each Genre

Genre	Average Tags/Video Without User Comments	Average Tags/Video With User Comments
Romance	1.45	1.55
Comedy	1.67	1.80
Horror	1.38	1.87
Sports	1.36	1.40
Tech	1.29	1.40
Average	**1.43**	**1.60**

Table 9: Average Predicted Tags/Video in Each genre

- Mostly a single tag and in certain cases bi-tags are assigned to the video
- Average number of tags/video increases with user comments
  - Greater contextual information available from user comments leading to genre overlap

Confusion Matrix

Genre	Romance	Comedy	Horror	Sports	Tech
Romance	80.16	**8.91**	3.23	4.45	3.64
Comedy	3.13	77.08	3.47	**9.03**	**7.29**
Horror	**10.03**	**9.34**	75.78	3.46	1.38
Sports	0.70	**7.30**	0	89.05	2.92
Tech	0.72	**5.07**	0.36	1.81	92.03

Table 10: Confusion matrix for Single Genre Prediction

Discussions : Confusion Matrix

- Romantic videos are frequently tagged as Comedy
  - Romantic movies or videos have light-hearted Comedy in them identifiable from the user comments
- Horror videos are frequently confused to be Comedy, as users frequently find them funny and not very scary
- Both Sports and Tech videos are sometimes tagged as Comedy
- Bias towards Comedy often arises out of the off-topic conversation between the users in the posts from the *jokes, teasing and* mostly **sarcastic comments** etc.
- Overall, from the precision figures, it seems that Sports and Tech videos are easy to distinguish from the remaining genres.

Comparison between Different Models

Model	Average F_1 Score
Multi-Class SVM Baseline: With User Comments	84.3606
Single Genre Prediction : Without User Comments + Without Wikipedia & WordNet	68.76
Single Genre Prediction : With User Comments + Without Wikipedia & WordNet	74.95
Single Genre Prediction : Without User Comments + With Wikipedia & WordNet	71.984
Single Genre Prediction : With User Comments+ With Wikipedia &WordNet	*80.9*
Multiple Genre Prediction : Without User Comments + With Wikipedia & WordNet	84.952
Multiple Genre Prediction : With User Comments + With Wikipedia & WordNet	**91.48**

Table 11: Average F_1-Score of Different Models

Issues

- Incorrect concept retrieval from Wikipedia due to ambiguous named entities
  - "*Manchester rocks*" can refer to Manchester United Football Club (Sports) or Manchester City (Place)
- Considering only WordNet synsets gives less coverage. Considering the gloss information helps to some extent.
  - It runs the risk of incorporating noise. Consider the word *good* and the gloss of one of its synsets {*dear, good, near -- with or in a close or intimate relationship*}.
  - "*good*" is associated to *Romance* due to "*relationship*"
- Uploader provided video meta-data is **typically small**
  - User comments provide information but incorporate noise as well
  - Auto-generated bot advertisements for products, off-topic conversation between users, fake urls and other spams
- Mis-spelt words, different forms of slangs and abbreviations mar the accuracy. Example : "*love*" spelt as "*luv*"

- THANK YOU