Discourse-based Sentiment Analysis

Parminder Bhatia **Pavan Kumar Ramnath**

1 Introduction

An essential phenomenon in natural language processing is the use of discourse relations to establish a coherent relation, linking phrases and clauses in a text. The presence of linguistic constructs like connectives, modals, conditionals and negation can alter sentiment at the sentence level as well as the clausal or phrasal level. Consider the example, ”@user share ’em! i’m quite excited about Tintin, despite not really liking original comics. Probably because Joe Cornish had a hand in.” The overall sentiment of this example is positive, although there is equal number of positive and negative words. This is due to the connective despite which gives more weight to the previous discourse segment. Any bag-of-words model would be unable to classify this sentence without considering the discourse marker. Consider another example, ”Think i’ll stay with the whole ’sci-fi’ shit. but this time...a classic movie.” The overall sentiment is again positive due to the connective but, which gives more weight to the following discourse segment. Thus it is of utmost importance to capture all these phenomena in a computational model.

In order to understand the problem, let us start with a simple example:
”Although I like the characters, the movie is horrible.”

For a human reader, the polarity of this sentence is clearly negative. However, in a classical (word-counting/bag of words) sentiment analysis approach, all feature words (like and horrible) would contribute equally to the total sentiment, thus yielding a verdict of a neutral or mixed polarity at best. This motivates us towards using Discourse Analysis for the Sentiment Analysis Problem. This could be accomplished by assigning different weights to distinct rhetorical roles, quantifying their contribution to the overall sentiment conveyed by a text .

2 Categorization of Discourse Relations

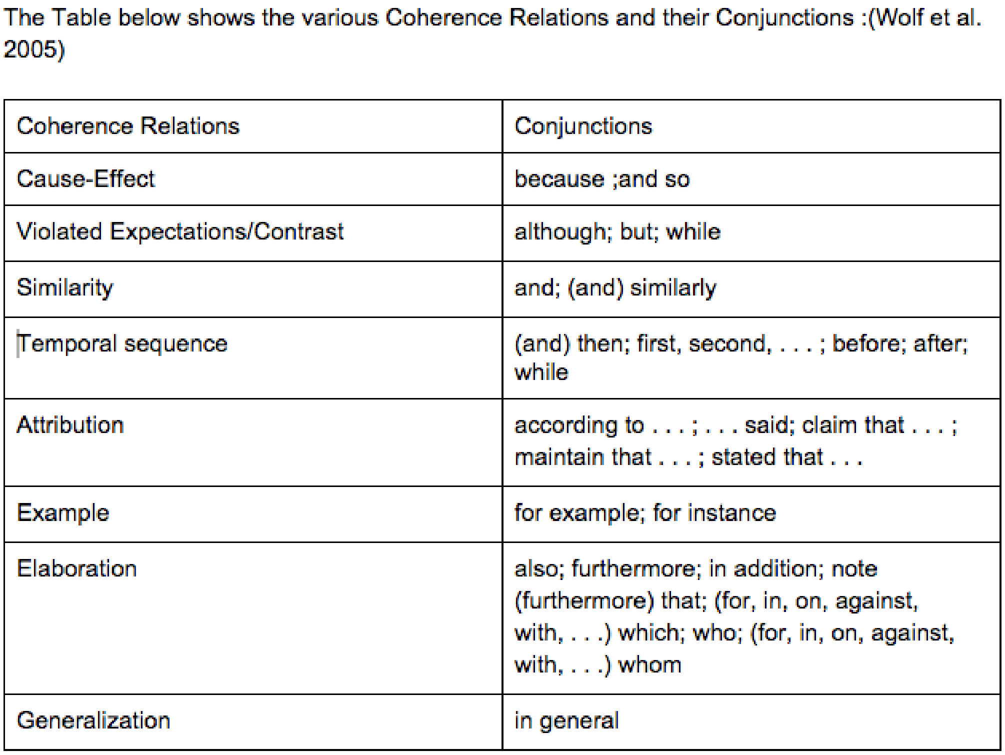
2.1 Discourse Parsing

Discourse Parsing deals with the identification of elementary discourse units (EDUs) and the relations between them. We shall use the RST parser from Ji and Eisenstein to mine for discourse relations which will in turn be used as inputs for feature extraction and sentiment analysis.

2.2 Discourse Relations for Sentiment Analysis

Listed in the figure below are key dicourse relations identified by Wolf and Gibson [2005]. Only some of these directly contribute to sentiment analysis and are considered below.

Violated Expectations/Contrast Consider the example - ”(I’m quite excited about Tintin), despite (not really liking original comics)”. A simple bag-of-words model will classify it as neutral. This is because it has one positive term excited and one negative phrase not really liking. However, it represents a positive emotion of the opinion holder, due to the segment after the connective despite. Violating expectation conjunctions oppose or refute the neighboring discourse segment.



The Table below shows the various Coherence Relations and their Conjunctions :(Wolf et al. 2005)

Coherence Relations	Conjunctions
Cause-Effect	because ;and so
Violated Expectations/Contrast	although; but; while
Similarity	and; (and) similarly
Temporal sequence	(and) then; first, second, . . . ; before; after; while
Attribution	according to . . . ; . . . said; claim that . . . ; maintain that . . . ; stated that . . .
Example	for example; for instance
Elaboration	also; furthermore; in addition; note (furthermore) that; (for, in, on, against, with, . . .) which; who; (for, in, on, against, with, . . .) whom
Generalization	in general

We further categorize them into the following two sub-categories: Conj-Fol and Conj-Prev. Conj-Fol is the set of conjunctions that give more importance to the discourse segment that follows them. Conj-Prev is the set of conjunctions that give more importance to the previous discourse segment.

Conclusive or Inferential Conjunctions These are the set of conjunctions, Conj-Infer, that tend to draw a conclusion or inference. Hence, the discourse segment following them (subsequently in Example 11) should be given more weight. Eg : I was nt much satisfied with ur so-called gud phone and subsequently decided to reject it.

Conditionals Conditionals tend to tone down polarity of an EDU. Consider for example the sentence below: ”If (MicroMax improved its battery life), (it wud hv been a gr8 product)”.

Both improve and gr8 represent a high degree of positive sentiment. But the presence of if tones down the final polarity as it introduces a hypothetical situation in the context. The if-then-else constructs depict situations which may or may not happen subject to certain conditions.

3 Discourse Relations

3.1 RST Parser

The RST parser is based on the enhancements mentioned in the paper Ji and Eisenstein by Ji and Eisenstein and is built from the framework of base features in it.

1 RST based Discourse Processing

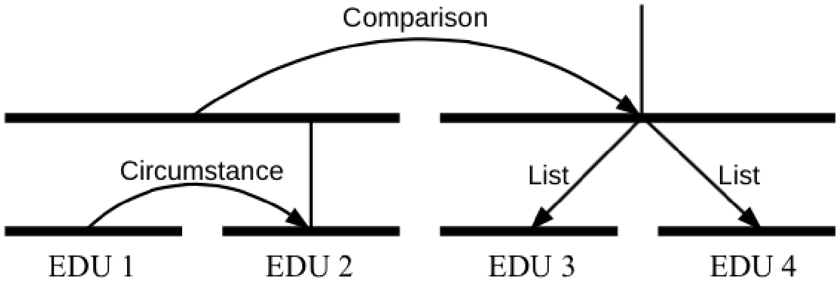


Figure 1: An example of RST tree.

Figure 1: An example of RST tree.

4 Feature Space/Algorithm

4.1 Background - RST Relation

- Generate EDUs: As RST Tree implementation requires the use of EDUs, the first part of the implementation involved creating EDUs for RST. For this step, we have implemented a parser, which is similar to the open source implementation of HILDA, for creating EDUs from the reviews.

- Creating Relation Tree: Once we have the EDUs, then we use the RST Tree implementation to generate the Tree Relation for the EDUs.

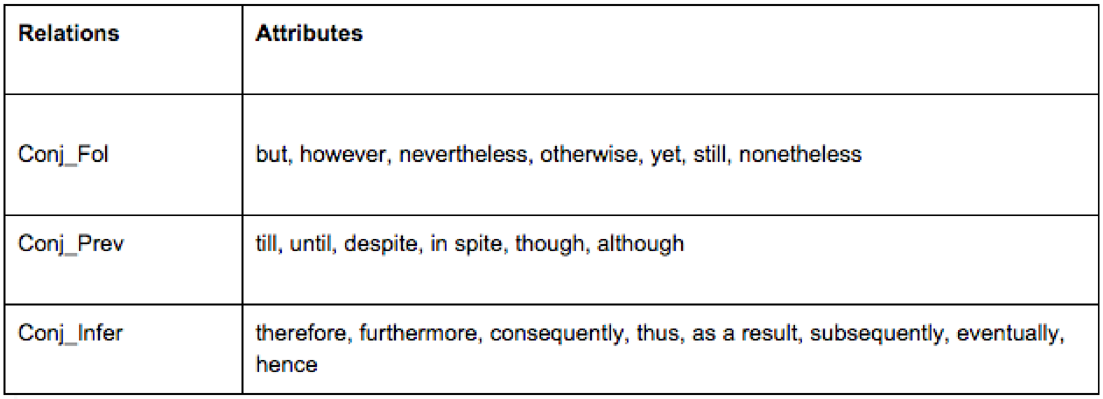
Using the above steps, we were able to generate the following relations:
elaboration-general-specific, elaboration-object-attribute-e, span, comment, attribution, interpretation-s, Sequence, Comparison, temporal-same-time, explanation-argumentative, concession, antithesis, evidence, circumstance-e, antithesis-e, attribution-n, attribution, attribution-e, Question-Answer, result, purpose, enablement, consequence-s, hypothetical, condition, Topic-Drift, restatement, Problem Solution, Contrast, Comparison, antithesis, antithesis-e, consequence-s, circumstance-e

4.2 Algorithm

The Support Vector Machines have been found to outperform other classifiers, like Nave Bayes and Maximum Entropy, in sentiment classification (Pang et al., 2002). Hence, in our work, SVMs are used to classify the set of feature vectors.

4.3 Feature Space

As base features in our implementation we use words post lemmatization and stemming to prevent the document-feature matrix from becoming too sparse. Apart from the base, the following discourse-based features are added:



Relations	Attributes
Conj_Fol	but, however, nevertheless, otherwise, yet, still, nonetheless
Conj_Prev	till, until, despite, in spite, though, although
Conj_Infer	therefore, furthermore, consequently, thus, as a result, subsequently, eventually, hence

Figure 2: List of feature words used in our analysis

1. **Relations:** As suggested in the section 2, we are interested in relations of type contrast, conclusion. With our current implementation, these relations are captured by **Constrast, Comparison, antithesis, antithesis-e, consequence-s, concession**

2. **Contrast:** As the table above suggests, we have identified the above relations, where Conj-Fol and Conj-Infer give more weight to the second part of structure as compared to the first one whereas Conj-Prev gives more weight to the first part.

   Thus if we find any of the above relation, we add feature of the type **(word ,"Contrast" ,1)** or **(word ,"Contrast" ,2)**. 1 and 2 are basically classes to categorize whether the word belongs to a more emphasis or less emphasis node.

3. **Height:** We also observed that height (between node and leaf-level) in the RST tree where the relation exists plays an important role in determining the final outcome. Thus, if we get a contrast relation at a very high level in a tree, then that relation would capture higher importance as compared to relation coming at the leaf level.

4. **Conditionals:** The presence of 'if' tones down the final polarity as it introduces a hypothetical situation in the context. The if-then-else constructs depict situations which may or may not happen subject to certain conditions. For this we basically evaluated **Problem-Solution** relation and if we found this relation then we would have to tone down weight of each of the words in the relation. For this we basically added feature of the type - **(word ,"Condition" ,3)** for each word in relation.

5. **Modals:** Similar to the case of conditionals , the modals (might, may, could, should, would etc.) depict irrealis events. The conditional does not necessarily talk of MicroMax being great, but talks of its possibility of being great subject to certain conditions (its battery life). These constructs cannot be handled by taking a simple majority valence of terms.

5 Evaluation and Results

For Evaluation we took 3 datasets.

- One was for movie reviews (Maas et al. [2011]) where we had about 2200 training data and about 1100 for testing. The ratings were in range from 1 to 10, where 1-4 were treated negative and 7 to 10 were treated as positive sentiments.
- Data set two was for Restaurant Reviews taken from Yelp Dataset [yel]. In this dataset we had about 2500 training and 1300 test reviews.

- Dataset 3 was also for Restaurant Reviews taken from Yelp Dataset [yel]. In this dataset we had about 2000 training and about 1000 test reviews.

5.1 Implementation

For implementation, we ran SVM first without using the discourse features and then by including the discourse features mentioned in the previous section.

Some of the results we obtained:

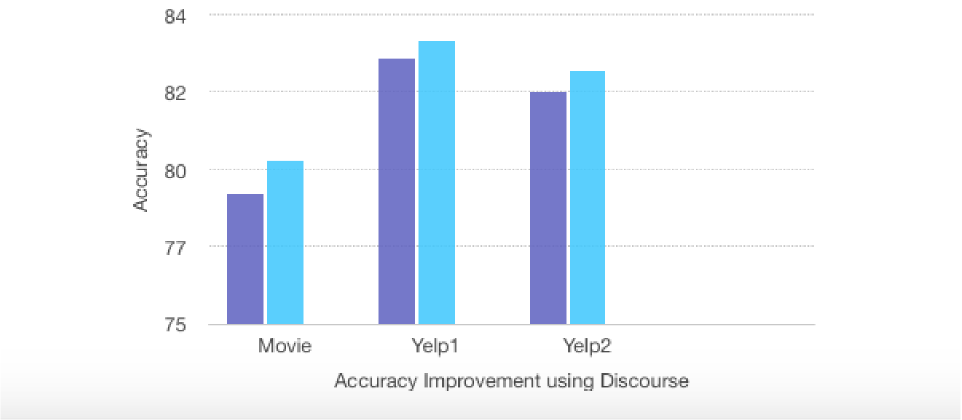


Figure 3: Accuracy

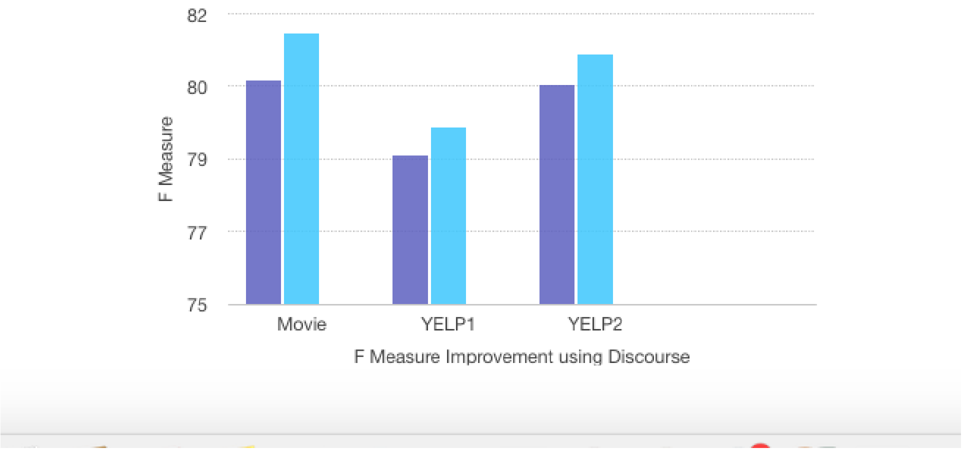


Figure 4: F1-Score

5.1.1 Dataset 1

For the movie reviews, the initial scores were **Accuracy - 0.787974683544, FScore - 0.803902439024**. When discourse features were added and trained, the results improved to **Accuracy - 0.79746835443 , FScore - 0.815384615385**.

5.1.2 Dataset 2

For the restaurant reviews (larger set), the initial scores were **Accuracy - 0.8270609319, FScore - 0.785793562708**. When the restaurant reviews were trained with discourse structures, the results improved to **Accuracy - 0.832437275986, FScore - 0.792452830189**.

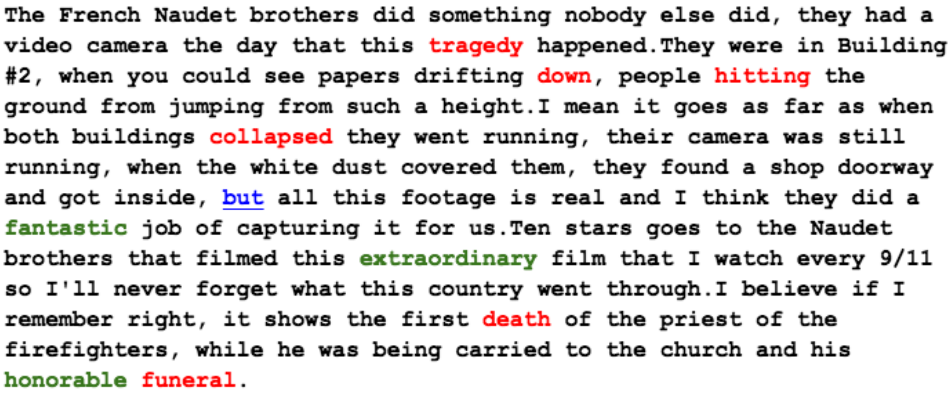
5.1.3 Dataset 3

For the restaurant reviews (smaller set), the initial scores were **Accuracy - 0.817725752508, FScore - 0.802893309222**. When the restaurant reviews were trained with discourse structures, the results imporved to **Accuracy - 0.823578595318 , FScore - 0.810081008101**.

5.2 Close Evaluation

As can be seen from above we are getting moderate improvements by the addition of these discourse features. To scrutinize further, we tried to evaluate the reviews which were earlier classified wrong but now are being classified correctly.

If you see, this paragraph has more negative words than positive hence and hence has been tagged as so by the words model. On the other hand, the discourse features from ”but all this footage is real and I think they did a fantastic job of capturing it for us” turns the tide especially as a very positive word such as fantastic is in the right side of a ’but’ constrast relation.



The French Naudet brothers did something nobody else did, they had a video camera the day that this tragedy happened.They were in Building #2, when you could see papers drifting down, people hitting the ground from jumping from such a height.I mean it goes as far as when both buildings collapsed they went running, their camera was still running, when the white dust covered them, they found a shop doorway and got inside, but all this footage is real and I think they did a fantastic job of capturing it for us.Ten stars goes to the Naudet brothers that filmed this extraordinary film that I watch every 9/11 so I'll never forget what this country went through.I believe if I remember right, it shows the first death of the priest of the firefighters, while he was being carried to the church and his honorable funeral.

Figure 5: Sample Getting Corrected with our Features

6 Discussion

Accuracy improvements over the baseline and the compared systems in all the datasets clearly signal the effectiveness of incorporating discourse information for sentiment classification. The bag-of-words model integrated with discourse information outperforms the bag-of-words model in all our settings, although the performance improvements vary in different settings.

6.1 Movie Reviews and Restaurant Reviews

We saw that movie reviews give better increase in performance compared to restaurant reviews. One of the reasons for this can be attributed to the fact that movie reviews in general are more elaborate and follow the sentence structure and relation more frequently as compared restaurant reviews which are generally written with lack of structure (sometimes similar to tweets).

7 Enhancements - Domain Adaptation/Cross-Domain Sentiment Classification

7.1 Introduction

Sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of users publishing sentiment data (e.g. reviews, blogs). Although traditional classification algorithms can be used to train sentiment classifiers from manually labeled text data, the labeling work can be time-consuming and expensive. Meanwhile, users often use some different words when they express sentiment in different domains. If we directly apply a classifier trained in one domain to other domains, the performance will be very low due to the differences between these domains.

7.2 State of the art algorithms

To bridge the gap between the domains, various methods have been proposed such a spectral feature alignment (SFA) algorithm to align domain-specific words from different domains into unified clusters, with the help of domain independent words as a bridge. In this way, the clusters can be used to reduce the gap between domain-specific words of the two domains, which can be used to train sentiment classifiers in the target domain accurately. State of art algorithms such as structural correspondence learning (SCL) algorithm exist on similar lines.

Domain Independent Features Firstly, we need to identify domain independent features. As mentioned above, domain-independent features should occur frequently and act similarly in both the source and target domains.

A first strategy is to select domain-independent features based on their frequency in both domains. A second strategy is based on the mutual dependence between features and labels on the source domain data. Mutual information is applied on source domain labeled data to select features as ”pivots”, which can be referred to as domain-independent features.

7.3 Algorithm - Discourse as Pivot Feature

As suggested in the image for SCL, finding the appropriate pivot features helps in generating a better cross domain classifier as can be seen by the difference in classification from the features generated by frequency as compared to Mutual Information. We believe that discourse features can also be treated as pivot features as they are to an extent independent of domain and dependent on the structure of the sentence, which is similar across domains.

To evaluate the performance in cross domain, let us consider 2 scenarios:

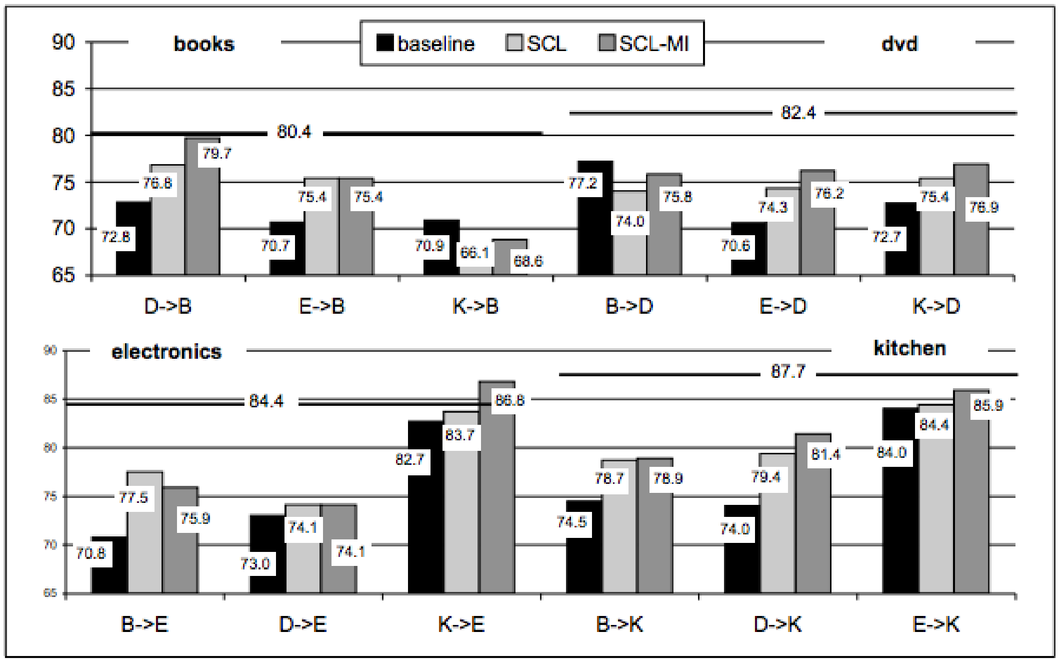


Figure 1: Accuracy results for domain adaptation between all pairs using SCL and SCL-MI. Thick black lines are the accuracies of in-domain classifiers.

Figure 6: Cross Domain Analysis using SCL

1. When we don't use discourse features and do cross domain sentiment analysis using SVM for training the data in source domain.
2. When we use discourse as feature and do cross domain sentiment analysis using SVM for training the data in source domain.

We shall train on source domain (with-without discourse) and observe its impact in target domain.

7.4 Evaluation for Domain Adaptation

For evaluation, we will use the same three datasets we used in the above parts (1 for Movie Reviews and 2 for Restaurant Reviews) to check their performance in cross-domain sentiment analysis.

Movie to Restaurant Reviews We have two datasets for restaurants which we shall call Y1 and Y2 and call Movie review as M. In this part, we trained the model on M dataset using SVM and evaluated its performance on Y1 and Y2.

- Y1 gave cross-domain classification results as **Accuracy - 0.636200716846, FScore - 0.672580645161** . When we trained movie reviews with discourse structures, it improved to **Accuracy - 0.64247311828, FScore - 0.677923076923**.
- Y2 gave cross classification results as **Accuracy - 0.680602006689, FScore - 0.740841248304**. When retrained with discourse structures, it improved to **Accuracy - 0.688782608696 , FScore - 0.748059742023**.

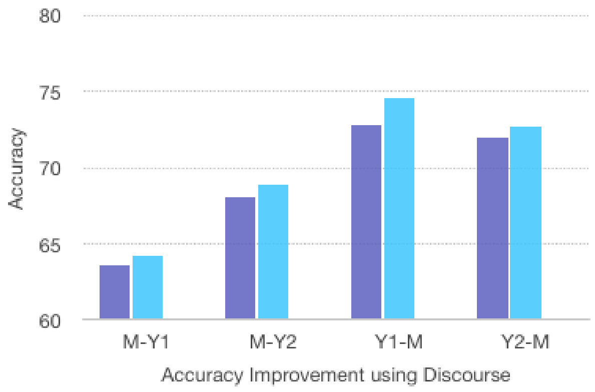


Figure 7: Accuracy

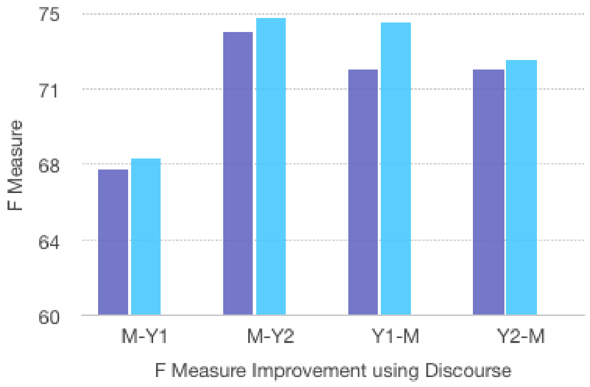


Figure 8: F1-Measure

Restaurant to Movie Reviews In this part, we trained the model on datasets Y1 and Y2 separately using SVM and evaluated its performance on M.

- When model was trained on Y1 without discourse features and evaluated on M, it gave cross-domain classification results as **Accuracy - 0.727848101266, FScore - 0.748538011696**. When we trained Y1 with discourse structures, it improved to **Accuracy - 0.745780590717, FScore - 0.766246362755**.
- When model was trained on Y2 without discourse features and evaluated on M, it gave cross-domain classification results as **Accuracy - 0.7194092827, FScore -**

0.722338204593. When we trained Y2 with discourse structures, it improved to **Accuracy - 0.726793248945, FScore - 0.72708113804**.

8 Conclusion

In this project we could perform a thorough analysis of how Discourse Features can be used for classification and their consistent superior performance in the three datasets tested on. We later see that these features can also play a crucial role in cross-domain classification, where they tend to improve the accuracy and f1-score and hence can be regarded as Domain Independent features.

References

URL "`https://www.yelp.com/dataset_challenge/dataset`".

Alexander Hogenboom Flavius Frasincar Uzay Kaymak Bas Heerschop, Frank Goossen and Franciska de Jong. Polarity analysis of texts using discourse structure. pages 1061–1070, 2011.

Heiner Stuckenschmidt Cacilia Zirn, Mathias Niepert and Michael Strube. Fine-grained sentiment analysis with structural features.

Vanessa Wei Feng and Graeme Hirst. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 60–68, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390524.2390534`.

Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing.

Alexander Hogenboom Jose M. Chenlo and David E. Losada. Sentiment-based ranking of blog posts using rhetorical structure theory.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. URL `http://www.aclweb.org/anthology/P11-1015`.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118693.1118704. URL `http://dx.doi.org/10.3115/1118693.1118704`.

Kenji Sagae. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing.

Pushpak Bhattacharyya "Subhabrata Mukherjee and Balamurali A. R.". "sentiment analysis in twitter with lightweight discourse analysis". 2012.

Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31, 2005.