TASS 2015, septiembre 2015, pp 99-104 recibido 20-07-15 revisado 24-07-15 aceptado 28-07-15

Sentiment Classification using Sociolinguistic Clusters

Clasificación de Sentimiento basada en Grupos Sociolingüísticos

Souneil Park
Telefonica Research
souneil.park@telefonica.com

Resumen: Estudios sociolingüísticos sugieren una alta similitud entre el lenguaje utilizado por personas de una misma clase social. Análisis recientes realizados a gran escala sobre textos en Internet y mediante el uso de mineria, sustentan esta hipótesis. Datos como la clase social del autor, su geolocalización o afinidades políticas tienen efecto sobre el uso del lenguaje en dichos textos. En nuestro trabajo utilizamos la información sociolingüística del autor para la identificación de patrones de expresión de sentimiento. Nuestro enfoque expande el ámbito del analisis de textos al análisis de los autores mediante el uso de su clase social y afinidad política. Más concretamente, agrupamos tweets de autores de clases sociales o afinidades políticas similares y entrenamos clasificadores de forma independiente con el propósito de aprender el estilo lingüístico de cada grupo. Este mismo enfoque podría mejorarse en combinación con otras técnicas de procesado del lenguaje y aprendizaje automático.
Palabras clave: sociolingüística, clase social, estilo lingüístico, clustering de usuario.

Abstract: Sociolinguistic studies suggest the similarity of language use among people with similar social state, and recent large-scale computational analyses of online text are providing various supports, for example, the effect of social class, geography, and political preference on the language use. We approach the tasks of TASS 2015 with sociolinguistic insights in order to capture the patterns in the expression of sentiment. Our approach expands the scope of analysis from the text itself to the authors: their social state and political preference. The tweets of authors with similar social state or political preference are grouped as a cluster, and classifiers are built separately for each cluster to learn the linguistic style of that particular cluster. The approach can be further improved by combining it with other language processing and machine learning techniques.
Keywords: Sociolinguistics, Social Group, Linguistic Styles, User Clustering.

1 Introduction

The social aspect of language is an important means for understanding commonalities and differences in the language use as communication is inherently a social activity. Shared ideas and preferences of people are reflected in the language use, and frequently observed from various linguistic features such as memes, style, and word choices. The social aspect is also clear in the expression of sentiment, especially in social media. The social media platforms have many elements that encourage the use of similar expressions among social groups. For example, retweets and hashtags facilitate the adoption of expressions, and the short length of messages encourages the use of familiar expressions.

Our approach to the tasks of TASS 2015 (Villena-Román et al., 2015) is based on the insights of sociolinguistics. Specifically, we focus on the effect of social variables on linguistic variations; people who share similar preference or status may show similarity in the expression of sentiment than others. For each task, we cluster the tweets by people who share some social features (e.g., political orientation, occupation, or football team preference). In order to capture the style of the sociolinguistic clusters, a classification model is trained separately for each cluster.

While the primary benefit of the approach is that it can distinguish the different style of

sentiment expression among different social groups, it also mitigates the scale limitation of the training data. For instance, some football players of the Social TV corpus and some entity-aspect pair of the STOMPOL corpus have limited number of associated tweets. Clustering them with other tweets that are spoken by people with similar preference expands the amount of data that can be used for training.

The approach can be easily combined with other language processing and machine learning techniques. Since our approach mainly considers the characteristics of the authors rather than the text of tweets itself, combining it with more advanced language processing techniques complements each other. In addition, there is much room for future improvement as the current implementation of our approach uses primitive language processing methods due to the limited local Spanish knowledge of the author.

2 Related Work

The increasing availability of large-scale text corpora and the advances of big data processing platforms allows computational analysis of sociolinguistic phenomena. Many works in NLP and computational social science nowadays are taking the hypotheses of socio-linguistics as well as other social sciences and testing them with online data sets.

In the context of computational analysis of sociolinguistics theories, a number of works showed the effect of social features on linguistic variations. For example, Eisenstein et al. (2011) observed the difference in term frequency depending on the demographics and geographical information of people, and also that the different language use can play a significant role in predicting the demographics of authors. A similar study was conducted with the information about occupation (Preotiuc-Pietro et al., 2015), and gender (Wang et al., 2013). There are also works that specifically observed the relation between the expression of sentiment and social variables, for example, daily routine (Dodds et al., 2011) and urban characteristics (Mitchell et al., 2013).

The difference of the language use depending on the political/ideological preference has been explored as well. In the communication literature, researchers have conceptualized the phenomena as *framing* (Scheufele, 1999) and many studies analyzed how political and social issues are framed differently between media outlets and partisan organizations, and how they are related with the perception of the public. Many works are applying computational methods for similar purposes and observing the difference of language use from various online text data, for example, news articles (Park et al., 2011a), comments (Park et al., 2011b), and discussion forums (Somasundaran et al., 2009).

3 System Design

The classification systems that we have developed for the tasks share the central idea of using sociolinguistic clusters. We describe below the system developed for each task in order.

The classification tool is kept identical for all the tasks. We use linear SVM equipped with the Elastic Net regularizer as the classifier. Given a set of tweets, the system trains a binary classifier for each class in a one-vs-all manner and combines them for multi-class classification. The input text of the classifier goes through the TFIDF bag of words transformation. We optionally applied lemmatization and stop-word removal with FreeLing (Carreras et al., 2004) to the system for Task 1.

3.1 Task 1: General Sentiment Classification

The corpus of this task includes the tweets of selected famous people and information about them. The information about the people includes the occupation and political orientation.

Our system for this task clusters people based on their information, and uses the tweets of the clusters for training. The idea behind the system is that people with the same occupation or political orientation will have similar patterns in the expression of sentiments. A similar idea was tested with English tweets in Preotiuc-Pietro's work (2015), where they predicted the occupation of authors based on

their tweets. For example, journalists may have a certain way of expressing the sentiment, which can be different from that of celebrities.

We tested various clustering of people: clustering by the occupation, political orientation, and by both occupation and political orientation. The system trains a classifier for each cluster, only using the tweets made by the people of that cluster. Depending on the task granularity (5-level or 3-level), the system trains the classifiers accordingly.

3.2 Task 2 (a): Aspect-based Sentiment Analysis with SocialTV corpus

Unlike Task 1, the corpus does not have the information about the authors; thus, it is not clear how to cluster the tweets. However, the unique characteristic of the topic (the football match between Real Madrid and F.C. Barcelona) and the aspect-sentiment pair of the tweets provide useful implications about the authors. The rivalry between the two teams suggests that many of the authors prefer one of the two, and the aspect-sentiment pair gives hints about the preferred team. For example, if a tweet discusses Xavier Hernández and its sentiment is positive, it is possible to guess that the author prefers F.C. Barcelona, and the author will share the sentiment with other fans of F.C. Barcelona, who will commonly share the sentiment towards either F.C. Barcelona or Real Madrid.

Thus, we group the aspects based on the team affiliation. The players of each team are grouped as a single entity respectively, and one classifier is developed for each team. The rest of the aspects (e.g., Afición) are not clustered since they do not share a common membership with either of the teams. Classifiers are also developed separately for the rest of the aspects.

3.3 Task 2 (b): Aspect-based Sentiment Analysis with STOMPOL corpus

For this task, we cluster tweets in two levels. First, we cluster tweets by the entity-aspect pair. Thus, even if the tweets cover the same entity (party), they are treated to cover a different topic if the covered aspect is not the same. For example, a tweet about the economic proposal (aspect) of Podemos (entity) is distinguished from a tweet about the education policy (aspect) of Podemos (entity). It is also possible to cluster tweets only by entity; however, we consider both elements for clustering as all the tweets of the corpus have a specific aspect in association to the entity. In addition, it is also frequent that people evaluate a political party in multiple ways regarding different aspects; a person may evaluate the economic policies of Podemos positively but negatively its foreign policies. Theories of political communication, such as agenda setting and framing theory, suggest that people often recognize the parties and issues together when they evaluate the parties.

Second, we further cluster the tweets based on the characteristics of the political parties. For example, following the left vs. right dimension, the tweets about the entity Izquierda Unida and the aspect Economia are grouped with those about Podemos and Economia as the two parties would have similarity in terms of economic policies than other parties on the right wing. As a result, 10 clusters are produced (2 party groups *x* 5 aspects) and a classifier is developed separately for each cluster.

We compared two ways of grouping of the parties: first is the left vs. right dimension as in the example, and the second is the new vs. old dimension considering the new political landscape of Spain. The detail of the party grouping is shown in Table 1.

Left vs. Right		Old vs. New	
PSOE	PP	PP	Podemos
Podemos	Cs	PSOE	UPyD
IU		IU	Cs
UPyD			

Table 1: Two groupings of the parties

4 Results and Discussion

4.1 Task 1 General Sentiment Classification (5-levels, Full corpus)

For this task, we ran three versions of the method; first, clustering of the authors by occupation, second, by political orientation, third, by both. We submitted the first version (cluster by occupation) as it performed better

than the other two. The performance metrics are summarized in Table 2. The result and the performance trend were similar for the 1k test set corpus so we only describe the result of the full-corpus.

	Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F-Measure
Occupation	0.462	0.385	0.450	0.415
Political Orientation	0.446	0.372	0.428	0.398
Both	0.423	0.351	0.401	0.374

Table 2. Performance Summary

The breakdown of the performance by sentiment category in Table 3 offers more insights. The performance for the category NEU and P is worse compared to that of other categories. While other optimization can be made for the two categories, we believe the method can be improved simply by having more number of examples of those categories in the training set. Compared to other categories, the current corpus includes much less examples for the two categories.

Category	P	R	F1
N	0.417	0.49	0.451
N+	0.35	0.539	0.424
NEU	0.064	0.217	0.099
NONE	0.659	0.405	0.502
P	0.094	0.554	0.161
P+	0.728	0.498	0.592

Table 3. Performance of Version 1 (Cluster-by-Occupation) by Sentiment Category

Interestingly, the performance further goes down when preprocessing (lemmatization and stopword removal) is conducted on the tweets. This performance drop was observed regardless of the version of our approach. The result suggests that conventional preprocessing removes important linguistic features that are relevant to sentiment expression. Due to the performance drop, we chose not to apply the preprocessing in the following tasks.

4.2 Task 1 General Sentiment Classification (3-levels, Full corpus)

We ran the same three versions of the method and the results are shown in Table 4. The performance is relatively higher than the 5-level classification task in general. Similar to the previous result, the version that clusters people by occupation performs better than the other two.

	Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F-Measure
Occupation	0.594	0.518	0.491	0.504
Political Orientation	0.566	0.469	0.476	0.472
Both	0.549	0.454	0.459	0.457

Table 4. Performance Summary

The performance breakdown shows some difference from the previous task. First of all, the performance for the category P is much higher. We believe this is because the number of training examples of this category is higher than the previous task; the examples of P+ and P categories are merged together. We also see similar improvement for the category N. The category NEU still remains as a bottleneck. The improvement observed in the categories N and P suggests that similar improvement may be achieved for the category NEU if there are more examples in the training set.

Category	P	R	F1
N	0.53	0.835	0.648
NEU	0.11	0.098	0.104
NONE	0.808	0.226	0.353
P	0.625	0.806	0.704

Table 5. Performance of Version 1 (Cluster-by-Occupation) by Sentiment Category

4.3 Task 2a Aspect-based Sentiment Analysis with SocialTV corpus

As described, the approach to this task is to group the tweets by aspects that share the team membership in the training phase. The

performance of the approach is shown in Table 6.

	Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F-Measure
Cluster-by-Team	0.631	0.460	0.484	0.471

Table 6. Performance Summary

Further analysis is required to understand the effect of the method. The breakdown of the performance by category does not show a clear pattern: while the tweets related to some players are identified very accurately but those of some other players are not; the performance does not differ much depending on the team of the players nor the sentiment expressed. We believe a larger test set that has enough samples for all players will better reveal the effect of the approach.

4.4 Task 2b Aspect-based Sentiment Analysis with STOMPOL corpus

Two versions of the approach are applied to the task: first, clustering the tweets of the same aspect by the parties of the same ideological leaning (left vs. right); second, by the novelty of the parties. The result is shown in Table 7.

	Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F-Measure
Left vs. Right	0.557	0.252	0.297	0.272
New vs. Old	0.521	0.250	0.280	0.264

Table 7. Performance Summary

The version that groups by the ideological leaning of the parties performed better than the other version. The breakdown of the performance revealed that the approach performed better for the tweets that express a negative sentiment in general. For example, nine categories out of the top-10 categories in terms of F1 score were those expressing a negative sentiment. This is partly because many tweets related to politics often convey a negative sentiment hence there are more training examples with the negative sentiment.

5 Conclusion

In this paper, we present a sentiment classification method that utilizes sociolinguistic insights. The method is based on the idea that people with similar social state (e.g., occupation) or political orientation may show similarity also in the way they express their sentiment online. Thus, the method is focused on grouping authors with similar taste or occupation. A classifier is developed separately for each group to capture the similarities and differences of expression particularly within the group.

The method achieves around 0.45 and 0.6 in terms of accuracy for the 5-level Task 1 classification and 3-level Task 1 classification, respectively. It achieves 0.63 and 0.56 for the Social TV corpus and for the STOMPOL corpus. The result shows that the method performs better for the sentiment classes with more training examples. It can also be further improved by combining it with more language processing methods optimized to Spanish.

References

Carreras, Xavier, Isaac Chao, Lluis Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proc. of LREC*.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS ONE*, 6(12), e26752.

Eisenstein, J, Noah A. S., and Eric P. X. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics,

and objective characteristics of place. *PLoS ONE* 8: e64417.

Park, S., Ko, M., Kim, J., Liu, Y., & Song, J. 2011a. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM conference on Computer supported cooperative work.*

Park, S., Lee, K., & Song, J. 2011b. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Preotiuc-Pietro, D., Lampos, V., & Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Scheufele, D. A. 1999. Framing as a theory of media effects. *Journal of communication*, 49(1), 103-122.

Somasundaran, S., & Wiebe, J. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Association for Computational Linguistics.*

Villena-Román, J., García-Morera, J., García-Cumbreras, M.A., Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A. 2015. Overview of TASS 2015. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*. CEUR-WS.org vol. 1397.

Wang, Y. C., Burke, M., Kraut, R. E. 2013. Gender, topic, and audience response: an analysis of user-generated content on facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.