Sentiment Translation for low resourced languages: Experiments on Irish General Election Tweets

Haithem Afli[1], Sorcha McGuire [2] and Andy Way[1]

ADAPT Centre
School of Computing,
Dublin City University,
Dublin, Ireland.
[1] {haithem.afli, andy.way}@adaptcentre.ie
[2] mcguirso@tcd.ie

Abstract. This paper presents two main methods of Sentiment Analysis (SA) of User-Generated Content for a low-resource language: Irish. The first method, automatic sentiment translation, applies existing English SA resources to both manually- and automatically-translated tweets. We obtained an accuracy of 70% using this approach. The second method involved the manual creation of an Irish-language sentiment lexicon: *SentiFoclóir*. This lexicon was used to build the first Irish SA system, *SentiFocalTweet*, which produced superior results to the first method, with an accuracy of 76%. This demonstrates that translation from Irish to English has a minor effect on the preservation of sentiment; it is also shown that the *SentiFocalTweet* system is a successful baseline system for Irish sentiment analysis.

Keywords: User-Generated Content, Social Media, Less-resourced languages, Sentiment translation.

1 Introduction

In this Golden Era of social media, tweeting on Twitter, posting on Facebook, or messaging on WhatsApp is a part of many people's daily routine. Social media provides an important platform for self-expression, free speech and debate for millions of users. With the rapid growth of such User-Generated Content (UGC) available on the Web, Sentiment Analysis (also known as Opinion Mining) can be applied to a number of different areas of interest to the general public [15].

In this paper, we focus on applying SA to the politics domain, by analysing tweets relating to the 2016 General Election in Ireland. The election discussion which took place on Twitter during that period constitutes a vast and vivid source of politics-inspired expressions of sentiment. More importantly, the participating Twitter users also happen to be real life voters.

While it is evident that most of the UGC produced globally these days is in English, the Twittersphere is very much a multicultural and multilingual space. The website Indigenous Tweets [1] records statistics on the use of 82 minority languages over Twitter; As of January 2017 the Irish language is the 6th most popular of those, with 1,501,225 tweets having been sent in Irish to date. Naturally, many of the tweets produced in relation to the Irish General Election were in Irish.

However, the majority of Sentiment Analysis (SA) research has been focused on English, with less attention paid to how it could be implemented for minority languages such as Irish. Unfortunately, Irish is also a low-resource language, with very limited linguistic data, corpora and tools currently in existence.

The aim of this paper is to explore how SA could be best implemented for low-resource languages, with a focus on Irish. In particular, we present what is – to the best of our knowledge – the first ever sentiment lexicon for the Irish language: *Senti-Foclóir*. We also focus on how existing Machine Translation and SA systems may be used for extracting sentiment from low-resource languages when the development of language-specific resources proves to be too time-consuming or costly. This is an approach which we call "Automatic Sentiment Translation".

2 Related Work

2.1 Sentiment Analysis

Sentiment Analysis (SA) is a discipline at the crossroads of computational linguistics which has received a lot of attention in academic research of late. SA is the computational study of opinions, sentiments and emotions as they are expressed in text.

The proliferation of opinion-rich data from UGC, which includes anything from political commentary to product reviews, has led to widespread opportunities for many practical applications of SA [9]. It is of huge interest to companies, for example, to be able to detect opinions on certain aspects of services and products as expressed in reviews and social media posts. In this domain, analysts from different fields are turning to the natural language processing community to sift through these kinds of texts and make sense of them [2]. [26] have shown that news and blog sentiment data are significantly correlated with some indicators in stock markets. SA was also used for some specific domains like Detecting Radical Content on Web Forums [22] or for Reviews Classification [4].

2.2 Different Methods of Sentiment Analysis

Unfortunately, as web language tends to be very informal and unstructured, littered with grammatical and spelling errors, hashtags and text speak, the sen-

[1] `http://indigenoustweets.com/`

timent analysis of social media text proves a very difficult task to undertake. A variety of techniques have been used to address this problem. These methods can be broadly classified into two main categories: approaches employing machine learning algorithms and approaches based on lexical resources and natural language processing.

Supervised and unsupervised machine learning methods have been proposed in the literature using different aspects of text as sources of features [1]. Prior studies on using several supervised learning algorithms for sentiment classification are seen in [18]. Their best performance was obtained using Support Vector Machines. [6]used a bootstrap parametric ensemble framework to develop their Twitter SA system.

Other studies [19, 23] used sentiment lexicons to build their SA systems.

Another distinction can be made between SA systems on the document-level [24, 17], sentence-level [7, 10] and phrase-level [3].

Almost all of the work on Twitter sentiment analysis can be linked to the sentence-level methods. But the informality and special characteristics of tweets (e.g. hashtags) lead to a great amount of noisy data, making tweet analysis considerably different from standard sentence-level tools [11]. Studies related to the sentiment analysis of tweets have focused mainly on: the development of unsupervised learning techniques [25], the integration of new features [20], collecting data and labeling techniques [3].

In this paper we are interested in using lexicon-based methods at word- and sentence-level applied to the task of Irish tweets classification. We present more details on this in Section 4.

2.3 Sentiment Translation

Previous exploration of the effect that translation has on the preservation of sentiment has been carried out on Arabic social media text [14, 21]. Their experiments revealed that automatic SA of English translations can produce competitive results when compared to those produced by state-of-the-art Arabic SA systems.

Our approach differs slightly in the sentiment ranges used; while the analysis approach used by [14] merely scored texts as being positive, negative or neutral, we have introduced a finer-grained scale, with posts classified for sentiment on a scale of 0 to 1, where: [0, 0.3] = high negative, [0.3, 0.5] = negative-neutral, [0.5, 0.7] = positive-neutral and [0.7, 1] = high positive.

3 The Irish Language

3.1 Current Status

Irish (*Gaeilge*) is a Celtic language; it enjoys constitutional status as the first official language of the Irish State and is also an official EU language. It was

widely spoken in Ireland up to the middle of the 18th century; however, after the Famine, it experienced a rapid decline in use. Recently, the language has experienced a revival, particularly among younger generations of Irish speakers based in urban areas such as Dublin. Nowadays, it is spoken, in varying degrees, by 1,774,437 people in the Republic of Ireland (2011 census), and is the first language of many inhabitants of the *Gaeltacht* (Irish-speaking) regions. There is also a great number of Irish language communities found overseas, among the Irish Diaspora and academics, in countries such as the US, Canada and Australia.

Interaction across this global linguistic community in the current digital age is greatly facilitated by the use of Irish language technologies and social media platforms. These channels of communication are a positive force in contributing to the preservation of this minority language and overcoming some of the obstacles it faces [8].

3.2 Linguistic Features of Irish

[8] outline certain linguistic features of the Irish language which pose a challenge for the development of language-processing technologies and SA systems. For example, Irish has no words for "yes" or "no". Instead, as in Latin, an affirmative/negative response is given by repeating the verb with or without a negative participle.

Irish is a morphologically rich language, and word forms are frequently inflected for semantic or grammatical reasons. Another feature which proves to be difficult for automatic processing to handle is the system of initial consonant mutations: the beginnings of words are changed through eclipsis (Irish: *séimhiú*) and lenition (*urú*) to satisfy various syntactic or morphological requirements. There are also two vowel-initial mutations: t-prosthesis and h-prosthesis. Figure 1 contains an example of *séimhiú*; 'gránna' (English: *"ugly"*) changes to 'ghránna' in certain conditions.

4 Irish Sentiment Analysis

4.1 *Senti-Foclóir* Lexicon

In this paper we present - to the best of our knowledge - the first ever Irish language sentiment lexicon: *Senti-Foclóir*[2]. This lexicon is based on the AFINN-111 lexicon [5], which was manually translated from English to Irish. *Senti-Foclóir* has been made freely available for download[3].

AFINN-111 contains 2,477 general English sentiment words and multi-word expressions (MWE); the corresponding *Senti-Foclóir* includes 5,571 words at

[2] *Foclóir* is the Irish word for "lexicon" or "dictionary"
[3] ADD LINK

present (due to the greater number of surface forms which exist for the majority of words in Irish). Efforts were made to include as many different surface forms of each term as possible, *e.g.* those modified by lenition and eclipsis.

The lexicon also contains the most common inflected forms of each verb listed. Most verbs appear in the lexicon in the following tenses: *past, past-passive, present, present-passive, future, future-passive.* In future work the lexicon may be extended to include other tenses, such as *conditional, subjunctive, imperative* etc.

Each word in the AFINN-111 lexicon is scored according to sentiment polarity with an integer between minus five (strongly negative) and plus five (strongly positive). This 11-point range allows for finer-granularity analysis; many of the current English language sentiment lexicons do not go beyond simply rating a word as *positive*, *negative* and *neutral*, which does not consider the degree of intensity of the sentiment expressed.

Listed in Figure 1 are some example entries from the *Senti-Foclóir* lexicon.

```
gránna  -3
ghránna -3
déistineach     -3
déistin -3
déistine        -3
cumas   2
cumais  2
gcumas  2
chumas  2
ar bord 1
as láthair      -1
neamhláithrí    -1
```

Fig. 1. Exanple entries from *Senti-Foclóir*

4.2 *SentiFocalTweet* System

Our Irish Sentiment Analysis system is an adaptation of an existing system implemented in the analysis of English language tweets related to the 2016 Irish General Election (Anon). The *SentiWordTweet* system uses a number of steps in the analysis, including: tokenization, part-of-speech tagging, word-sense disambiguation, and word scoring. It consults a number of general sentiment lexicons (such as *SentiWordNet* [4] and AFINN-111), as well as domain-specific lexicons containing vocabulary related to Irish politics.

[4] SentiWordNet is an opinion lexicon derived from the *WordNet* [13] database.

***GA* Tweet**	@trevorocSF @sinnfeinireland Maith sibh! Tuiscint maith agaibh ar saincheisteanna daoine óga #GE16 #MakeASmartVote https://t.co/HJdRwIk9VN
GA(Manl.Sent)	0.9
GA(Auto.Sent)	0.832
EN(Manl.Trans.)	@trevorocSF @sinnfeinireland Well done! You have a good understanding of young people's issues #GE16 #MakeASmartVote https://t.co/HJdRwIk9VN
EN(Manl.Trans., Manl.Sent)	0.8
EN(Manl.Trans., Auto.Sent)	0.806
EN(Auto.Trans.)	@sinnfeinireland @trevorocSF Well done! You good impression on young people's issues # GE16 #MakeASmartVote https://t.co/HJdRwIk9VN
EN(Auto.Trans., Manl.Sent)	0.9
EN(Auto.Trans., Auto.Sent)	0.858

Table 1. An example of Irish tweet with corresponding translations and sentiment scores.

The *SentiFocalTweet* baseline system architecture is more elementary, as it does not implement POS-tagging or word-sense disambiguation. After tokenisation and sentence-splitting, each word from the tweet is searched for within the *Senti-Foclóir* lexicon and given the corresponding score. If the word is not found in the lexicon, it taken to be neutral. The final score assigned to the whole tweet is the sum of all the positive and negative word scores; the system then outputs a sentiment score between 0 (negative) and 1 (positive).

Match Pair	**% Match**
a. *GA(Manl.Sent) - GA(Auto.Sent)*	76%
b. *GA(Manl.Sent) - EN(Manl.Trans., Manl.Sent)*	95.5%
c. *GA(Manl.Sent) - EN(Manl.Trans., Auto.Sent)*	70%
d. *GA(Manl.Sent) - EN(Auto.Trans., Manl.Sent)*	80%
e. *GA(Manl.Sent) - EN(Auto.Trans., Auto.Sent)*	70%
f. *EN(Manl.Trans., Manl.Sent) - EN(Auto.Trans., Manl.Sent)*	74%
g. *EN(Manl.Trans., Manl.Sent) - EN(Manl.Trans., Auto.Sent)*	69.5%
h. *EN(Auto.Trans., Manl.Sent) - EN(Auto.Trans., Auto.Sent)*	66.5%
i. *EN(Auto.Trans., Auto.Sent) - EN(Manl.Trans., Manl.Sent)*	65%

Table 2. Percentage match between pairs of sentiment scores for Irish tweets.

Match Pair	% Match
a. *EN(Manl.Sent) - EN(Auto.Sent)*	50%
b. *EN(Manl.Sent) - GA(Manl.Trans., Manl.Sent)*	81%
c. *EN(Manl.Sent) - GA(Manl.Trans., Auto.Sent)*	46%
d. *EN(Manl.Sent) - GA(Auto.Trans., Manl.Sent)*	61.5%
e. *EN(Manl.Sent) - GA(Auto.Trans., Auto.Sent)*	42%
f. *GA(Manl.Trans., Manl.Sent) - GA(Auto.Trans., Manl.Sent)*	69.5%
g. *GA(Manl.Trans., Manl.Sent) - GA(Manl.Trans., Auto.Sent)*	61.5%
h. *GA(Auto.Trans., Manl.Sent) - GA(Auto.Trans., Auto.Sent)*	73.5%
i. *GA(Auto.Trans., Auto.Sent) - GA(Manl.Trans., Manl.Sent)*	55.5%

Table 3. Percentage match between pairs of sentiment scores for English tweets.

5 Experimental Results

5.1 Sentiment Scores

For the evaluation of the different sentiment scoring methods and our Irish *SentiFocalTweet* system we prepared two separate test sets: 200 English (*EN*) and 200 Irish (*GA*) tweets related to the 2016 General Election. These tweets were scored manually for sentiment by an Irish Twitter user with good knowledge of the Irish political landscape. These scores will be referred to as *EN(Manl.Sent)* and *GA(Manl.Sent)* respectively, in line with the naming conventions used by [14].

From these two test sets, a number of other sentiment scores were generated. The *SentiWordTweet* and *SentiFocalTweet* systems were used to analyse the original tweets and output automatic sentiment scores for each, which we have called *EN(Auto.Sent)* and *GA(Auto.Sent)*.

Manual translations of the original tweets were obtained using crowd-sourcing; we created a website to allow members of the Irish language community to contribute to the research by providing translations in either direction, EN to GA or GA to EN. The tweets were also automatically translated using Google Translate[5], a widely-used general domain Statistical Machine Translation system. The resulting *GA(Manl.Trans)*, *EN(Manl.Trans)*, *GA(Auto.Trans.)* and *EN(Auto.Trans.)* corpora were scored both manually and automatically for sentiment.

An example tweet, along with its corresponding translations and sentiment scores, is seen in Table 1.

5.2 Evaluation

Tables 2 and 3 indicate the percentage agreement between the various sentiment score datasets for both GA to EN and EN to GA, respectively. Each pair of scores was manually marked for agreement; two scores were said to be in agreement if they fell between the same range. The evaluation ranges used were: [0,

[5] `https://translate.google.ie/`

0.3] (high negative), [0.3, 0.5] (negative-neutral), [0.5, 0.7] (positive-neutral) and [0.7, 1] (high positive). For example, row a. in Table 2 shows the percentage match between the manual scores and the *SentiFocalTweet* system scores for the original GA tweets corpus. The accuracy of the Irish SA system using the *Senti-Foclóir* lexicon was found to be 76%.

Row e. is also of particular interest; it shows the percentage match between the original manual score and the *SentiWordTweet* automatic score after Machine Translation had been applied from GA to EN. This method, with an accuracy of 70%, performs slightly worse than the Irish SA system. This demonstrates that there is a slight loss in the preservation of the original sentiment during translation into English, but it does not have a drastic impact.

Looking at the results for the EN corpus in Table 3, it is apparent that both methods did not perform as well on English tweets. After a comparison of the EN and GA test sets, it appears that the language used in Irish tweets is far more clear cut and less noisy, with fewer instances of text speak, sarcasm, etc. It is possible that this is why a disparity in score accuracy has arisen with the English tweets dataset.

6 Conclusion

We have presented two main methods of sentiment analysis for a low-resource language: Irish. Our evaluation shows that automatic sentiment analysis of GA to EN translated tweets produces competitive results when compared to the gold-standard manual scoring of the original tweet. We also present a baseline sentiment analysis system for Irish, which implements *Senti-Foclóir*, the first sentiment lexicon created in Irish. This system outperforms the translation-into-English method of sentiment analysis, with an increase in accuracy of 6%. These results suggest that when there is a lack of lexical resources available in the focus language, it is possible to use existing English resources and Machine Translation to achieve a relatively accurate sentiment analysis. However, this paper also indicates that creating and using a language-specific sentiment lexicon can produce superior results w.r.t. the automatic sentiment translation method.

The *SentiFocalTweet* baseline system presented in this paper may be improved in future work through the implementation of an Irish stemmer[6], Twitter-specific part-of-speech tagging [12] and word-sense disambiguation using the recently-developed WordNet Gaeilge [16].

[6] `https://gist.github.com/kscanne/2048178`

References

1. Ahmed Abbasi, Ammar Hassan, and Milan Dhar. Benchmarking twitter sentiment analysis tools. In *Proceedings of the 9th Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.
2. Joseph Davies-Gavin, Clarence Lee, and Lingling Zhang. The rise of big data analytics for marketing. In *Marketing Science Institute Conference on Big Data*, Massachusetts Institute of Technology, Cambridge, USA, 2012.
3. Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford University, 2009.
4. Alaa Hamouda and Mohamed Rohaim. Reviews classification using sentiwordnet lexicon. In *The Online Journal on Computer Science and Information Technology (OJCSIT)*, W11-0123, 2011.
5. Lars Kai Hansen, Adam Arvidsson, Finn Arup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news - affect and virality in twitter. In *The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011)*, 2011.
6. Asif Hassan, Ali Abbasi, and Deze Zeng. Twitter sentiment analysis: A bootstrap ensemble framework. In *Proceedings of the ASE/IEEE International Conference on Social Computing*, pages 357–364, Washington D.C., USA, 2013.
7. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington, USA, 2004.
8. John Judge, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P Scannell, and Elaine Uí Dhonnchadha. *The Irish language in the digital age*. Springer, 2012.
9. Siavash Kazemian, Shunan Zhao, and Gerald Penn. Evaluating sentiment analysis evaluation: A case study in securities trading. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 119–127, Baltimore, Maryland, USA, 2014.
10. Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the COLING conference*, pages 1367–1373, Geneva, Switzerland, 2004.
11. Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 538–541, Barcelona, Catalonia, Spain, 2011.
12. Teresa Lynn, Kevin Scannell, and Eimear Maguire. Minority language twitter: Part-of-speech tagging and analysis of irish tweets. *ACL-IJCNLP 2015*, page 1, 2015.
13. George A. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, 1995.
14. Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 1:1–20, 2015.
15. Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC USA, 2010.
16. Jim O'Regan, Kevin Scannell, and Elaine Uı Dhonnchadha. lemongawn: Wordnet gaeilge as linked data. In *LDL 2016 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, page 36, 2016.
17. Bo Pang and Lillian Lee. *Opinion Mining and Sentiment Analysis*, 2008.

18. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, Jeju Island, Korea, 2002.
19. Horacio Saggion and Adam Funk. Interpreting sentiwordnet for opinion classification. In *Proceedings of the 7th Conference on Language Resources and Evaluation*, Malta 2010.
20. Hassan Saif, Yulan He, and Harith Alani. Alleviating data sparsity for twitter sentiment analysis. In *The 2nd Workshop on Making Sense of Microposts*, Lyon, France, 2012.
21. Mohammad Salameh, Saif M Mohammad, and Svetlana Kiritchenko. Sentiment after translation: A case-study on arabic social media posts. In *In proceedings of The 2015 Annual Conference of the North American Chapter of the ACL*, pages 767–777, 2015.
22. Chalothorn Tawunrat and Ellman Jeremy. Using sentiwordnet and sentiment analysis for detecting radical content on web forums. In *Proceedings of 6th Conference on Software, Knowledge, Information Management and Applications (SKIMA 2012)*, Chengdu University, China, 2012.
23. Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. In *Journal of the American Society for Information Science and Technology*, volume 61(12), pages 2544–2558, 2010.
24. Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA, 2002.
25. Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of CIKM'11*, pages 1031–1040, Glasgow, Scotland, UK 2011.
26. Wenbin Zhang and Steven Skiena. Trading strategies to exploit blog and news sentiment. In *The 4th International AAAI Conference on Weblogs and Social Media*, 2010.