"What is Your Evidence?" A Study of Controversial Topics on Social Media

Aseel A. Addawood **Masooda N. Bashir**
School of Information Sciences
University of Illinois at Urbana-Champaign
Aaddaw2, mnb@illinois.edu

Abstract

In recent years, social media has revolutionized how people communicate and share information. One function of social media, besides connecting with friends, is sharing opinions with others. Micro blogging sites, like Twitter, have often provided an online forum for social activism. When users debate about controversial topics on social media, they typically share different types of evidence to support their claims. Classifying these types of evidence can provide an estimate for how adequately the arguments have been supported. We first introduce a manually built gold standard dataset of 3000 tweets related to the recent FBI and Apple encryption debate. We develop a framework for automatically classifying six evidence types typically used on Twitter to discuss the debate. Our findings show that a Support Vector Machine (SVM) classifier trained with n-gram and additional features is capable of capturing the different forms of representing evidence on Twitter, and exhibits significant improvements over the unigram baseline, achieving a F_1 macro-averaged of 82.8%.

1 Introduction

Social media has grown dramatically over the last decade. Researchers have now turned to social media, via online posts, as a source of information to explain many aspects of the human experience (Gruzd & Goertzen, 2013). Due to the textual nature of online users' self-disclosure of their opinions and views, social media platforms present a unique opportunity for further analysis of shared content and how controversial topics are argued.

On social media sites, especially on Twitter, user text contains arguments with inappropriate or missing justifications—a rhetorical habit we do not usually encounter in professional writing. One way to handle such faulty arguments is to simply disregard them and focus on extracting arguments containing proper support (Villalba and Saint-Dizier, 2012; Cabrio and Villata, 2012). However, sometimes what seems like missing evidence is actually just an unfamiliar or different type of evidence. Thus, recognizing the appropriate type of evidence can be useful in assessing the viability of users' supporting information, and in turn, the strength of their whole argument.

One difficulty of processing social media text is the fact that it is written in an informal format. It does not follow any guidelines or rules for the expression of opinions. This has led to many messages containing improper syntax or spelling, which presents a significant challenge to attempts at extracting meaning from social media content. Nonetheless, we believe processing such corpora is of great importance to the argumentation-mining field of study. Therefore, the motivation for this study is to facilitate online users' search for information concerning controversial topics. Social media users are often faced with information overload about any given topic, and understanding positions and arguments in online debates can potentially help users formulate stronger opinions on controversial issues and foster personal and group decision-making (Freeley and Steinberg, 2013).

Continuous growth of online data has led to large amounts of information becoming available for others to explore and understand. Several automatic techniques have allowed us to determine different viewpoints expressed in social media text, e.g., sentiment analysis and opinion mining. However, these techniques struggle to identify complex relationships between concepts in the text. Analyzing argumentation from a computational linguistics point of view has led very recently to a new field called argumentation mining (Green et al., 2014).

Proceedings of the 3rd Workshop on Argument Mining, pages 1–11,
Berlin, Germany, August 7-12, 2016.

It formulates how humans disagree, debate, and form a consensus. This new field focuses on identifying and extracting argumentative structures in documents. This type of approach and the reasoning it supports is used widely in the fields of logic, AI, and text processing (Mochales and Ieven, 2009). The general consensus among researchers is that an argument is defined as containing a claim, which is a statement of the position for which the claimant is arguing. The claim is supported with premises that function as evidence to support the claim, which then appears as a conclusion or a proposition (Walton, Reed, & Macagno, 2008; Toulmin, 2003).

One of the major obstacles in developing argumentation mining techniques is the shortage of high-quality annotated data. An important source of data for applying argumentation techniques is the web, particularly social media. Online newspapers, blogs, product reviews, etc. provide a heterogeneous and growing flow of information where arguments can be analyzed. To date, much of the argumentation mining research has been limited and has focused on specific domains such as news articles, parliamentary records, journal articles, and legal documents (Ashley and Walker, 2013; Hachey and Grover, 2005; Reed and Rowe, 2004). Only a few studies have explored arguments on social media, a relatively under-investigated domain. Some examples of social media platforms that have been subjected to argumentation mining include Amazon online product reviews (Wyner, Schneider, Atkinson, & Bench-Capon, 2012) and tweets related to local riot events (Llewellyn, Grover, Oberlander, & Klein, 2014).

In this study, we describe a novel and unique benchmark data set achieved through a simple argument model, and elaborate on the associated annotation process. Unlike the classical Toulmin model (Toulmin, 2003), we search for a simple and robust argument structure comprising only two components: a claim and associated supporting evidence. Previous research has shown that a claim can be supported using different types of evidence (Rieke and Sillars, 1984). The annotation that is proposed in this paper is based on the type of evidence one uses to support a particular position on a given debate. We identify six types, which are detailed in the methods section (Section 3). To demonstrate these types, we collected data regarding the recent Apple/FBI encryption debate on Twitter between January 1 and March 31, 2016. We believe that understanding online users' views on this topic will help scholars, law enforcement officials, technologists, and policy makers gain a better understanding of online users' views about encryption.

In the remainder of the paper, Section 2 discusses survey-related work, Section 3 describes the data and corresponding features, Section 4 presents the experimental results, and Section 5 concludes the paper and proposes future directions.

2 Related Work

2.1 Argumentation mining

Argumentation mining is the study of identifying the argument structure of a given text. Argumentation mining has two phases. The first consists of argument annotations and the second consists of argumentation analysis. Many studies have focused on the first phase of annotating argumentative discourse. Reed and Rowe (2004) presented Araucaria, a tool for argumentation diagramming that supports both convergent and linked arguments, missing premises (enthymemes), and refutations. They also released the AracuariaDB corpus, which has been used for experiments in the argumentation mining field. Similarly, Schneider et al. (2013) annotated Wikipedia talk pages about deletion using Walton's 17 schemes (Walton 2008). Rosenthal and McKeown (2012) annotated opinionated claims, in which the author expresses a belief they think should be adopted by others. Two annotators labeled sentences as claims without any context. Habernal, Eckle-Kohler & Gurevych (2014) developed another well-annotated corpus,to model arguments following a variant of the Toulmin model. This dataset includes 990 instances of web documents collected from blogs, forums, and news outlets, 524 of which are labeled as argumentative. A final smaller corpus of 345 examples was annotated with finer-grained tags. No experimental results were reported on this corpus.

As far as the second phase, Stab and Gurevych (2014b) classified argumentative sentences into four categories (none, major claim, claim, premise) using their previously annotated corpus (Stab and Gurevych 2014a) and reached a 0.72 macro-F1

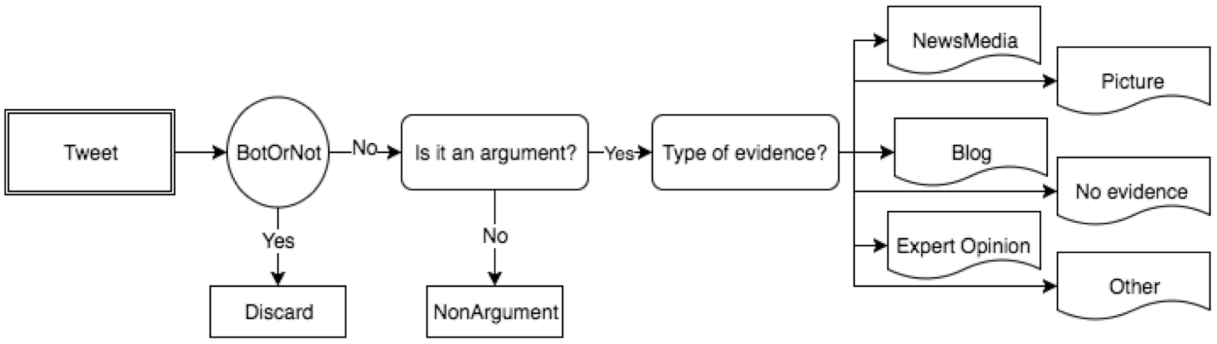

Figure 1: flow chart for annotation

score. Park and Cardie (2014) classified propositions into three classes (unverifiable, verifiable non-experimental, and verifiable experimental) and ignored non-argumentative text. Using multi-class SVM and a wide range of features (n-grams, POS, sentiment clue words, tense, person) they achieved a 0.69 Macro F1.

The IBM Haifa Research Group (Rinott et al., 2015) developed something similar to our research; they developed a data set using plain text in Wikipedia pages. The purpose of this corpus was to collect context-dependent claims and evidence, where the latter refers to facts (i.e., premises) that are relevant to a given topic. They classified evidence into three types (study, expert, anecdotal). Our work is different in that it includes more diverse types of evidence that reflect social media trends while the IBM Group's study was limited to looking into plain text in Wikipedia pages.

2.2 Social Media As A Data Source For Argumentation Mining

As stated previously there are only a few studies that have used social media data as a source for argumentation mining. Llewellyn et al. (2014) experimented with classifying tweets into several argumentative categories, specifically claims and counter-claims (with and without evidence), and used verification inquiries previously annotated by Procter, Vis, and Voss (2013). They used unigrams, punctuations, and POS as features in three classifiers. Schneider and Wyner (2012) focused on online product reviews and developed a number of argumentation schemes—inspired by Walton et al. (2008)—based on manual inspection of their corpus.

By identifying the most popular types of evidence used in social media, specifically on Twitter, our research differs from the previously mentioned studies because we are providing a social media annotated corpus. Moreover, the annotation is based on the different types of premises and evidence used frequently in social media settings.

3 Data

This study uses Twitter as its main source of data. Crimson Hexagon (Etlinger & Amand, 2012), a public social media analytics company, was used to collect every pubic post from January 1, 2016 through March 31, 2016. Crimson Hexagon houses all public Twitter data going back to 2009. The search criterion for this study was searching for a tweet that contains the word "encryption" anywhere in its text. The sample only included tweets from accounts that set English as their language; this was filtered in when requesting the data. However, some users set their account language to English, but constructed some tweets in a different language. Thus, forty accounts were removed manually, leaving 531,593 tweets in our dataset.

Although most Twitter accounts are managed by humans, there are other accounts managed by automated agents called social bots or Sybil accounts. These accounts do not represent real human opinions. In order to ensure that tweets from such accounts did not enter our data set, in the annotation procedure, we ran each Twitter user through the Truthy BotOrNot algorithm (Davis et al., 2016). This cleaned the data further and excluded any user with a 50% or greater probability of being a bot. Overall, 946 (24%) bot accounts were removed.

4 Methodology

4.1 Coding Scheme

In order to perform argument extraction from a social media platform, we followed a two-step approach. The first step was to identify sentences containing an argument. The second step was to identify the evidence-type found in the tweets classified as argumentative. These two steps were performed in conjunction with each other. Annotators were asked to annotate each tweet as either having an argument or not having an argument. Then they were instructed to annotate a tweet based on the type of evidence used in the tweet. Figure 1 shows the flow of annotation.

After considerable observation of the data, a draft-coding scheme was developed for the most used types of evidence. In order to verify the applicability and accuracy of the draft-coding scheme, two annotators conducted an initial trial on 50 randomized tweets to test the coding scheme. After some adjustments were made to the scheme, a second trial was conducted consisting of 25 randomized tweets that two different annotators annotated. The resulting analysis and discussion led to a final revision of the coding scheme and modification of the associated documentation (annotation guideline). After finalizing the annotation scheme, two annotators annotated a new set of 3000 tweets. The tweets were coded into one of the following evidence types.

News media account (*NEWS*) refers to sharing a story from any news media account. Since Twitter does not allow tweets to have more than 140 characters, users tend to communicate their opinions by sharing links to other resources. Twitter users will post links from official news accounts to share breaking news or stories posted online and add their own opinions. For example:

Please who don't understand encryption or technology should not be allow to legislate it. There should be a test... https://t.co/I5zkvK9sZf

Expert opinion (*EXPERT*) refers to sharing someone else's opinion about the debate, specifically someone who has more experience and knowledge of the topic than the user. The example below shows a tweet that shares a quotation from a security expert.

RT @ItIsAMovement "Without strong encryption, you will be spied on systematically by lots of people" - Whitfield Diffie

Blog post (*BLOG*) refers to the use of a link to a blog post reacting to the debate. The example below shows a tweet with a link to a blog post. In this tweet, the user is sharing sharing a link to her own blog post.

I care about #encryption and you should too. Learn more about how it works from @Mozilla at https://t.co/RTFiuTQXyQ

Picture (*PICTURE*) refers to a user sharing a picture related to the debate that may or may not support his/her point of view. For example, the tweet below shows a post containing the picture shown in figure 2.

RT @ErrataRob No, morons, if encryption were being used, you'd find the messages, but you wouldn't be able to read them

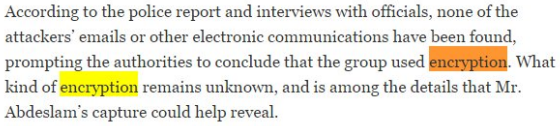

Figure 2: an example of sharing a picture as evidence

Other (*OTHER*) refers to other types of evidence that do not fall under the previous annotation categories. Even though we observed Twitter data in order to categorize different, discrete types of evidence, we were also expecting to discover new types while annotating. Some new types we found while annotating include audio, books, campaigns, petitions, codes, slides, other social media references, and text files.

No evidence (*NO EVIDENCE*) refers to users sharing their opinions about the debate without having any evidence to support their claim. The example below shows an argumentative tweet from a user who is in favor of encryption. However, he/she does not provide any evidence for his/her stance.

I hope people ban encryption. Then all their money and CC's can be stolen and they'll feel better knowing terrorists can't keep secrets.

Non Argument (*NONARG*) refers to a tweet that does not contain an argument. For example, the following tweet asks a question instead of presenting an argument.

RT @cissp_googling what does encryption look like

Another NONARG situation is when a user shares a link to a news article without posting any opinions about it. For example, the following tweet does not present an argument or share an opinion about the debate; it only shares the title of the news article, "Tech giants back Apple against FBI's 'dangerous' encryption demand," and a link to the article.

Tech giants back Apple against FBI's 'dangerous' encryption demand #encryption https://t.co/4CUushsVmW

Retweets are also considered NONARG because simply selecting "retweet" does not take enough effort to be considered an argument. Moreover, just because a user retweets something does not mean we know exactly how they feel about it; they could agree with it, or they could just think it was interesting and want to share it with their followers. The only exception would be if a user retweeted something that was very clearly an opinion or argument. For example, someone retweeting Edward Snowden speaking out against encryption backdoors would be marked as an argument. By contrast, a user retweeting a CNN news story about Apple and the FBI would be marked as NONARG.

Annotation discussion. While annotating the data, we observed other types of evidence that did not appear in the last section. We assumed users would use these types of evidence in argumentation. However, we found that users mostly use these types in a non-argumentative manner, namely as a means forwarding information. The first such evidence type was "scientific paper," which refers to sharing a link to scientific research that was published in a conference or a journal. Here is an example:

A Worldwide Survey of Encryption Products. By Bruce Schneier, Kathleen Seidel & Saranya Vijayakumar #Cryptography https://t.co/wmAuvu6oUb

The second such evidence type was "video," which refers to a user sharing a link to a video related to the debate. For example, the tweet below is a post with a link to a video explaining encryption.

An explanation of how a 2048-bit RSA encryption key is created https://t.co/JjBWym3poh

4.2 Annotation results

The results of the annotation are shown in Table 1 and Table 2. The inter-coder reliability was 18% and 26% for the two tasks, respectively, yielding a 70% inter-annotator observed agreement for both tasks. The unweighted Cohen's Kappa score was 0.67 and 0.79, respectively, for the two tasks.

Argumentation classification	Class distribution
Argument (ARG)	1,271
Non argument (NONARG)	1,729
Total	3000

Table 1: Argumentation classification distribution over tweets

Evidence type	Class distribution
No evidence	630
News media accounts	318
Blog post	293
Picture	12
Expert opinion	11
Other	7
Total	1,271

Table 2: Evidence type distribution over tweets

5 Experimental Evaluation

We developed an approach to classify tweets into each of the six major types of evidence used in Twitter arguments.

5.1 Preprocessing

Due to the character limit, Twitter users tend to use colloquialisms, slang, and abbreviations in their tweets. They also often make spelling and grammar errors in their posts. Before discussing feature selection, we will briefly discuss how we compensated for these issues in data preprocessing. We first replaced all abbreviations with their proper word or phrase counterparts (e.g., 2night => tonight) and replaced repeated characters with a single character (e.g., haaaapy => happy). In addition, we lowercased all letters (e.g., ENCRYPTION => encryption), and removed all URLs and mentions to other users after initially recording these features.

5.2 Features

We propose a set of features to characterize each type of evidence in our collection. Some of these features are specific to the Twitter platform. However, others are more generic and could be applied to other forums of argumentation. Many features follow previous work (Castillo, Mendoza, & Poblete, 2011; Agichtein, Castillo, Donato, Gionis, & Mishne, 2008). The full list of features appears in appendix A. Below, we identify four types of features based on their scope: Basic, Psychometric, Linguistic, and Twitter-specific.

Basic Features refer to N-gram features, which rely on the word count (TF) for each given unigram or bigram that appears in the tweet.

Psychometric Features refer to dictionary-based features. They are derived from the linguistic enquiry and word count (LIWC). LIWC is a text analysis software originally developed within the context of Pennebaker's work on emotional writing (Pennebaker & Francis, 1996; Pennebaker, 1997). LIWC produces statistics on eighty-one different text features in five categories. These include psychological processes such as emotional and social cognition, and personal concerns such as occupational, financial, or medical worries. In addition, they include personal core drives and needs such as power and achievement.

Linguistic Features encompass four types of features. The first is grammatical features, which refer to percentages of words that are pronouns, articles, prepositions, verbs, adverbs, and other parts of speech or punctuation. The second type is LIWC summary variables. The newest version of LIWC includes four new summary variables (analytical thinking, clout, authenticity, and emotional tone), which resemble "person-type" or personality measures.

The LIWC webpage ("Interpreting LIWC Output", 2016) describes the four summary variables as follows. *Analytical thinking* "captures the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns." *Clout* "refers to the relative social status, confidence, or leadership that people display through their writing or talking." *Authenticity* "is when people reveal themselves in an authentic or honest way," usually by becoming "more personal, humble, and vulnerable." Lastly, with *emotional tone*, "although LIWC includes both positive emotion and negative emotion dimensions, the tone variable puts the two dimensions into a single summary variable."

The third type is sentiment features. We first experimented with the Wilson, Wiebe & Hoffmann (2005) subjectivity clue lexicon to identify sentiment features. However, we decided to use the sentiment labels provided by the LIWC sentiment lexicon. We found it provides more accurate results than we would have had otherwise. For the final type, subjectivity features, we did use the Wilson et al. (2005) subjectivity clue lexicon to identify the subjectivity type of tweets.

Twitter-Specific Features refer to characteristics unique to the Twitter platform, such as the length of a message and whether the text contains exclamation points or question marks. In addition, these features encompass the number of followers, number of people followed ("friends" on Twitter), and the number of tweets the user has authored in the past. Also included is the presence or not of URLs, mentions of other users, hashtags, and official account verification. We also considered a binary feature for tweets that share a URL as well as the title of the URL shared (i.e., the article title).

6 Experimental results

Our first goal was to determine whether a tweet contains an argument. We used a binary classification task in which each tweet was classified as either argumentative or not argumentative. Some previous research skipped this step (Feng and Hirst, 2011), while others used different types of classifiers to achieve a high level of accuracy (Reed and Moens, 2008; Palau and Moens, 2009).

In this study, we chose to classify tweets as either containing an argument or not. Our results confirm previous research showing that users do not frequently utilize Twitter as a debating platform (Smith, Zhu, Lerman & Kozareva, 2013). Most individuals use Twitter as a venue to spread information instead of using it as a platform through which to have conversations about controversial issues. People seem to be more interested in spreading information and links to webpages than in debating issues.

As a first step, we compared classifiers that have frequently been used in related work: Naïve Bayes

Feature Set	Decision tree			SVM			NB		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1	Pre.	Rec.	F_1
UNI (*Base*)	72.5	69.4	66.3	81	78.5	77.3	69.7	67.3	63.9
All features	87.3	87.3	87.2	89.2	89.2	**89.2**	79.3	79.3	84.7

Table 3: Summary of the argument classification results in %

(NB) approaches as used in Teufel and Moens (2002), Support Vector Machines (SVM) as used in Liakata et al. (2012), and Decision Trees (J48) as used in Castillo, Mendoza, & Poblete (2011). We used the Weka data mining software as used in Hall et al. (2009) for all approaches.

Before training, all features were ranked according to their information gain observed in the training set. Features with information gain less than zero were excluded. All results were subject to 10-fold cross-validation. Since, for the most part, our data sets were unbalanced, we used the ''Synthetic Minority Oversampling TEchnique'' (SMOTE) approach (Chawla, Bowyer, Hall & Kegelmeyer, 2002). SMOTE is one of the most renowned approaches to solve the problem of unbalanced data. Its main function is to create new minority class examples by interpolating several minority class instances that lie together. After that, we randomized the data to overcome the problem of overfitting the training data.

Argument classification Regarding our first goal of classifying tweets as argumentative or non-argumentative, Table 3 shows a summary of the classification results. The best overall performance was achieved using SVM, which resulted in a 89.2% F_1 score for all features compared to basic features, unigram model. We can see there is a significant improvement from just using the baseline model.

Evidence type classification our second goal was for evidence type classification, results across the training techniques were comparable; the best results were again achieved by using SVM, which resulted in a 78.6% F1 score. Table 4 shows a summary of the classification results. The best overall performance was achieved by combining all features.

In table 5, we computed Precision, Recall, and F1 scores with respect to the top-used three evidence types, employing one-vs-all classification problems for evaluation purposes. We chose the top-used evidence types since other types were too small and could have led to biased sample data. The results show that the SVM classifier achieved a F_1 macro-averaged score of 82.8%. As the table shows, the baseline outperformed Linguistic and Psychometric features. This was not expected. However, Basic features (N-gram) had very comparable results to those from combining all features. In other words, the combined features captured the characteristics of each class. This shows

Feature Set	Decision tree			SVM			NB		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1	Pre.	Rec.	F_1
UNI (*Base*)	59.1	61.1	56.3	63.7	62.1	56.5	27.8	31.6	19.4
All features	76.8	77	76.9	78.5	79.5	**78.6**	62.4	59.4	52.5

Table 4: Summary of the evidence type classification results in %

Feature Set	NEWS vs. All			BLOG vs. All			NO EVIDENCE vs. All			Macro Average F1
	Pre.	Rec.	F_1	Pre.	Rec.	F_1	Pre.	Rec.	F_1	
UNI (*Base*)	76.8	74	73.9	67.3	64.4	63.5	78.5	68.7	65.6	67.6
Basic Features	842	81.3	81.3	85.2	83	82.9	80.1	75.5	74.4	79.5
Psychometric Features	62	61.7	57.9	64.6	63.7	63.5	59.2	58.9	58.6	60
Linguistic Features	65	65.3	64.2	69.1	69	69	63.1	62.6	62.4	65.2
Twitter-Specific Features	65.7	65.2	65	63.7	63.6	63.6	68.7	68.1	67.9	65.5
All features	84.4	84	**84.1**	86	85.2	**85.2**	79.3	79.3	**79.3**	**82.8**

Table 5: Summary of evidence type classification results using one-vs-all in %

that we can distinguish between classes using a concise set of features with equal performances.

6.1 Feature Analysis

The most informative features for the evidence type classification are shown in Table 6. There are different features that work for each class. For example, Twitter-specific features such as title, word count, and WPS are good indicators of the NEWS evidence type. One explanation for this is that people often include the title of a news article in the tweet with the URL, thereby engaging the aforementioned Twitter-specific features more fully.

Another example is that linguistic features like grammar and sentiments are essential for using the BLOG evidence type. The word "wrote," especially, appears often to refer to someone else's writing, as in the case of a blog. The use of the BLOG evidence type also seemed to correlate with emotional tone and negative emotions, which is a combination of positive and negative sentiment. This may suggest that users have strong negative opinions toward blog posts.

Feature Set	All Features
NEWS vs. All	Word count, title, personal pronoun, common adverbs, WPS, "iphone", "nsa director"
BLOG vs. All	Emotional Tone, 1st person singular, negation, colon, conjunction, "wrote", negative emotions, "blog"
NO EVIDENCE vs. All	Title,1st person singular, colon, Impersonal pronouns, discrepancies, insight, differentiation (cognitive processes), period, adverb, positive emotion

Table 6: Most informative features for combined features for evidence type classification

Concerning the NO EVIDENCE type, a combination of linguistic features and psychometric features best describe the classification type. Furthermore, in contrast with blogs, users not using any evidence tend to express more positive emotions. That may imply that they are more confident about their opinions. There are, however, mutual features used in both BLOG and NO EVIDENCE types as 1st person singular and colon. One explanation for this is that since blog posts are often written in a less formal, less evidence-based manner than news articles, they are comparable to tweets that lack sufficient argumentative support. One further shared feature is that "title" appears frequently in both NEWS and NO EVIDENCE types. One explanation for this is that "title" has a high positive value in NEWS, which often involves highlighting the title of an article, while it has a high negative value in NO EVIDENCE since this type does not contain any titles of articles.

As Table 5 shows, "all features" outperforms other stand-alone features and "basic features," although "basic features" has a better performance than the other features. Table 7 shows the most informative feature for the argumentation classification task using the combined features and unigram features. We can see that first person singular is the strongest indication of arguments on Twitter, since the easiest way for users to express their opinions is by saying "I …".

Feature set	Features
Unigram	I'm, surveillance, love, I've, I'd, privacy, I'll, hope, wait, obama
All	1st person singular, RT, personal pronouns, URL, function words, user mention, followers, auxiliary verbs, verb, analytic

Table 7: Most informative features argumentation classification

7 Conclusions and future work

In this paper, we have presented a novel task for automatically classifying argumentation on social media for users discussing controversial topics like the recent FBI and Apple encryption debate. We classified six types of evidence people use in their tweets to support their arguments. This classification can help predict how arguments are supported. We have built a gold standard data set of 3000 tweets from the recent encryption debate. We find that Support Vector Machines (SVM) classifiers trained with n-grams and other features capture the different types of evidence used in social media and demonstrate significant improvement over the unigram baseline, achieving a macro-averaged F1 score of 82.8 %.

One consideration for future work is classifying the stance of tweets by using machine learning techniques to understand a user's viewpoint and opinions about a debate. Another consideration for

future work is to explore other evidence types that may not be presented in our data.

Acknowledgments

We would like to thank Andrew Marturano and Amy Chen for assisting us in the annotation process.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 183-194). ACM.

Ashley, K. D., & Walker, V. R. (2013). From InformaNon Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study.

Cabrio, E., & Villata, S. (2012, July). Combining textual entailment and argumentation theory for supporting online debates interactions. In*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 208-212). Association for Computational Linguistics.

Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675-684). ACM.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 321-357.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016, April). BotOrNot: A System to Evaluate Social Bots. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 273-274). International World Wide Web Conferences Steering Committee.

Etlinger, S., & Amand, W. (2012, February). Crimson Hexagon [Program documentation]. Retrieved April, 2016, from http://www.crimsonhexagon.com/wp-content/uploads/2012/02/CrimsonHexagon_Altimeter_Webinar_111611.pdf

Feng, V. W., & Hirst, G. (2011, June). Classifying arguments by scheme. In*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 987-996). Association for Computational Linguistics.

Freeley, A., & Steinberg, D. (2013). *Argumentation and debate*. Cengage Learning.

Green, N., Ashley, K., Litman, D., Reed, C., & Walker, V. (2014). *Proceedings of the First Workshop on Argumentation Mining*. Baltimore, MD: Association for Computational Linguistics.

Gruzd, A., & Goertzen, M. (2013, January). Wired academia: Why social science scholars are using social media. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 3332-3341). IEEE.

Habernal, I., Eckle-Kohler, J., & Gurevych, I. (2014, July). Argumentation Mining on the Web from Information Seeking Perspective. In *ArgNLP*.

Hachey, B., & Grover, C. (2005, March). Sequence modelling for sentence classification in a legal summarisation system. In *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 292-296). ACM.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

Interpreting LIWC Output. Retrieved April 17, 2016, from http://liwc.wpengine.com/interpreting-liwc-output/

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, *28*(7), 991-1000.

Llewellyn, C., Grover, C., Oberlander, J., & Klein, E. (2014). Re-using an Argument Corpus to Aid in the Curation of Social Media Collections. In *LREC*(pp. 462-468).

Mochales, R., & Ieven, A. (2009, June). Creating an argumentation corpus: do theories apply to real arguments?: a case study on the legal argumentation of the ECHR. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law* (pp. 21-30). ACM.

Palau, R. M., & Moens, M. F. (2009, June). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law* (pp. 98-107). ACM.

Park, J., & Cardie, C. (2014, June). Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining* (pp. 29-38).

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological science*, *8*(3), 162-166.

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion*, *10*(6), 601-626.

Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, *16*(3), 197-214.

Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, *13*(04), 961-979.

Rieke, R. D., & Sillars, M. O. (1984). *Argumentation and the decision making process*. Addison-Wesley Longman.

Rinott, R., Dankin, L., Alzate, C., Khapra, M. M., Aharoni, E., & Slonim, N. (2015, September). Show Me Your Evidence–an Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in NLP (EMNLP), Lisbon, Portugal* (pp. 17-21).

Rosenthal, S., & McKeown, K. (2012, September). Detecting opinionated claims in online discussions. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on* (pp. 30-37). IEEE.

Schneider, J., & Wyner, A. (2012). Identifying Consumers' Arguments in Text. In *SWAIE* (pp. 31-42).

Schneider, J., Samp, K., Passant, A., & Decker, S. (2013, February). Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1069-1080). ACM.

Smith, L. M., Zhu, L., Lerman, K., & Kozareva, Z. (2013, September). The role of social media in the discussion of controversial topics. In *Social Computing (SocialCom), 2013 International Conference on* (pp. 236-243). IEEE.

Stab, C., & Gurevych, I. (2014a). Annotating Argument Components and Relations in Persuasive Essays. In *COLING* (pp. 1501-1510).

Stab, C., & Gurevych, I. (2014b). Identifying Argumentative Discourse Structures in Persuasive Essays. In *EMNLP* (pp. 46-56).

Teufel, S., & Kan, M. Y. (2011). *Robust argumentative zoning for sensemaking in scholarly documents* (pp. 154-170). Springer Berlin Heidelberg.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, *28*(4), 409-445.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Villalba, M. P. G., & Saint-Dizier, P. (2012). Some Facets of Argument Mining for Opinion Analysis. *COMMA*, *245*, 23-34.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347-354). Association for Computational Linguistics.

Wyner, A., Schneider, J., Atkinson, K., & Bench-Capon, T. J. (2012). Semi-Automated Argumentative Analysis of Online Product Reviews. *COMMA*,*245*, 43-50.

Appendix A. Feature types used in our Model

Type	Feature	Description
Basic Features	Unigram	Word count for each single word that appears in the tweet
	Bigram	Word count for each two words that appears in the tweet
Psychometrics Features	Perceptual process	Percentage of words that refers to multiple sensory and perceptual dimensions associated with the five senses.
	Biological process	Percentage of words related to body, health, sexual and Ingestion
	Core Drives and Needs	Percentage of words related to personal drives as power, achievement, reward and risk
	Cognitive Process-es	Percentage of words related to causation, discrepancy, tentative, certainty, inhi-bition and inclusive.
	Personal Concerns	Percentage of words related to work, leisure, money, death, home and religion
	Social Words	Percentage of words that are related to family and friends
Linguistic Features	Analytical Think-ing	Percentage of words that captures the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns
	Clout	Percentage of words related to the relative social status, confidence, or leader-ship that people display through there writing or talking.
	Authenticity	Percentage of words that reveals people in an authentic or honest way, they are more personal, humble, and vulnerable
	Emotional Tone	Percentage of words related to the emotional tone of the writer which is a com-bination of both positive emotion and negative emotion dimensions.
	Informal Speech	Percentage of words related to informal language markers as assents, fillers and swears words
	Time Orientation	Percentage of words that refer to Past focus, present focus and future focus.
	Grammatical	Percentage of words that refer to personal pronouns, impersonal pronouns, arti-cles, prepositions, auxiliary verbs, common adverbs, punctuation
	Positive emotion	Percentage of positive words in a sentence
	Negative emotion	Percentage of negative words in a sentence
	Subjectivity type	Subjectivity type derived by Wilson et al. (2005) lexicon
	Punctuation	Percentage of punctuation in text including periods, commas, colons, semico-lons etc.
Twitter-specific Features	RT	1.0 if the tweet is a retweet
	Title	1.0 if the tweet contains a title to the article title
	Mention	1.0 if the tweet contains a mention to another user '@'
	Verified account	1.0 if the author has a 'verified' account
	URL	1.0 if the tweet contains a link to a URL
	Followers	Number of people this author is following at posting time
	Following	Number of people following this author at posting time
	Posts	Total number of user's posts
	hashtag	1.0 if the tweet contains a hashtag '#'
	WC	Word count of the tweet
	Words>6 letters	Count of words with more then six letters
	WPS	Count of words per sentence
	QMark	Percentage of words contains question mark
	Exclam	Percentage of words contains exclamation mark