# Wheats the Deal? Understanding the GMO debate in online forums

Arman Irani, Kevin M. Esterling, Michalis Faloutsos, Deborah Pagliaccia
University of California, Riverside
{airan002, kevin.esterling, deborahp}@ucr.edu and michalis@cs.ucr.edu

***Abstract*—How can we comprehensively understand the main concerns and beliefs of the GMO debate in online forums? Genetically Modified Organisms (GMOs) have historically been a hotly debated topic, both within and outside of the agriculture industry. Understanding the complexity of these beliefs can lend policy makers the knowledge necessary to counteract misinformation. In this paper we develop *Forumlyze*, a systematic framework to understand user beliefs in online discourse surrounding an issue. As a case study, we focus on data collected from Reddit between 2019-2020 from four sub-forums: farming, agriculture, horticulture, and vegetable gardening. In our approach we (a) illustrate the fundamental and temporal characteristics of the issue (b) extract and characterize sentiments surrounding the issue (c) uncover the dominate concepts prevalent in this discussion and the context surrounding these concepts. The comprehensive nature of this analysis led to the following results. (1) The dominant concepts surrounding GMOs are Climate Change, Monsanto and Soil Science. (2) The sentiment of discourse around GMOs and its related concepts indicates a polarized affective system. (3) Evidence that real-world events impact online forum communities' sentiment surrounding GMOs-related concepts.**

**Keywords: Online forum mining, GMOs, Agriculture**

## I. Introduction

*How can we comprehensively understand the prevailing discourse for a specific issue in online forums?* This is the question we address in this work.

Online forums are an invaluable resource to the study of topical societal issues and concerns. They provide an accessible space for anyone to express opinions, in a vast and generally unconstrained manner. Furthermore, online forums often provide the ability for users to form their own communities disjoint from one another, with each pertaining to a niche topic of interest. These "micro-communities" create valuable text data because they are unique social networks dispensing qualitative emotional and personal insight into subject matter that polls or surveys are unable to capture.

Genetically Modified Organisms (GMOs) have been increasingly used over the past decade, with a simultaneous increase in the debate of their effect on humans and society. This work presents a framework to aid in the discussion of how to reach and educate in a era of digital information.

We discuss previous works and how this work relates to it in section IV.

**Contribution:** We propose *Forumlyze*, a systematic framework to understand controversial issues in online forums in depth. Our analysis consists of three levels: (a) we start with an **issue** of interest (e.g. GMO), for which we identify; (b) related **concepts** (e.g. Monsanto); and (c) the context for these concepts (e.g. cancer). Our approach consists of methods for identifying, quantifying, and characterizing posts of varying sentiment in regards to a pre-defined issue. From a technical point of view, we create a comprehensive suite of techniques by adopting and customizing state-of-the-art approaches: (a) aspect-based sentiment analysis. (b) BERTopic [1], a state-of-the-art pre-trained BERT-based topic modeling technique, and (c) Non-Negative Matrix Factorization to extract coherent narratives from our clusters of posts.

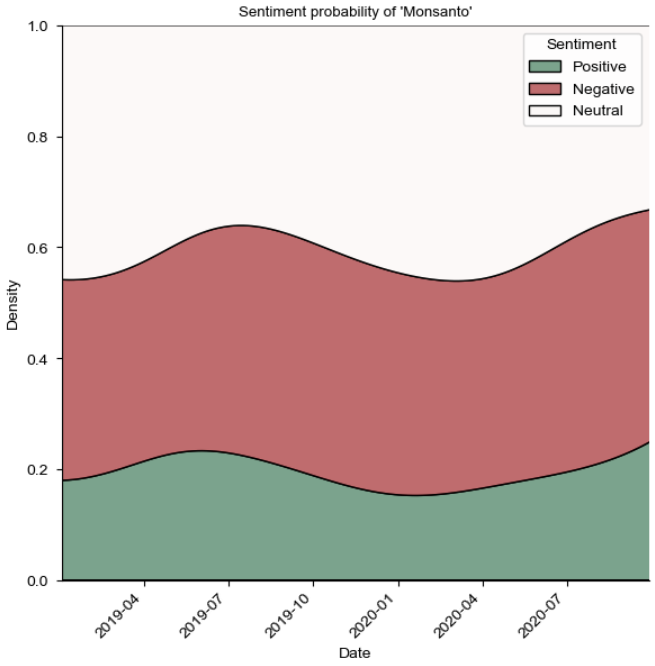


**Fig. 1:** Temporal sentiment probability density for the concept 'Monsanto', which we found to be a related concept in the GMO discussion in agricultural-focused online forums.

Using our suite of techniques, we conduct a real world study using 205,000+ posts and 35,000+ unique users from the online forum platform Reddit between January 2019 to September 2020. First we investigate the proliferation of the GMO debate within these communities. Next, we systematically identify and explore the main concepts surrounding GMOs. Finally, we reveal the narratives driving the concepts being disseminated through the communities.

The key observations are captured in the following points:

**a. The most popular concepts surrounding GMOs are Climate Change, Monsanto, and Soil Science.** Generalizing a multi-layered issue like GMOs to just one dimension dimin-

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

| Forum | # Threads | # Posts | # Users |
|---|---|---|---|
| r/vegetablegardening | 19,249 | 115,914 | 16,995+ |
| r/farming | 12,441 | 69,683 | 11,198+ |
| r/agriculture | 4,525 | 5,876 | 3,035+ |
| r/horticulture | 2,792 | 14,521 | 3,822+ |

**TABLE I:** Reddit dataset description.

ishes the complexity of the topic. In order to understand the negativity or positivity we must reveal the concepts being discussed within the context of GMOs. Our systematic approach reveals these three as being the most prevalent and important in the debate.

**b. All the concepts are polarized in sentiment.** We discover a proliferation of polarization from the highest level down to the lowest. Concepts are classified as maintaining relatively equal levels of Negative, Neutral, and Positive sentiment indicating continued polarized perceptions and attitudes towards GMOs.

**c. There is evidence of direct relationships between real-world events and online sentiment.** There exists sentiment volatility during various periods of time, and investigation into events occurring at similar times indicates empirical correlation between concept-level sentiment shifts and real-world events related to the concepts. This is important in understanding the impact the real-world has on user perception in online forums.

## II. Datasets and Methodology

**1. Reddit.** We selected these four subreddit communities because they accurately represent the diversity of the agricultural industry. r/VegetableGardening contain hobbyist's who engage in conversations regarding small-scale and personal gardening. r/Horticulture is a science-centric community discussing cultivation and farming. r/Farming is a community primarily of large-scale farmers who specialize in monoculture (barley, wheat, corn, soy). r/Agriculture focuses on the discussion of livestock and crops for human consumption. These communities were all active and contained a reasonable amount of unique users participating. To collect this data, we use the archiver service Pushshift [2] which collects every post made on Reddit and makes the data publicly available for academic purposes.

**2. Sentiment detection model.** We use the customized aspect-based sentiment analysis (ABSA) introduced in RAFFMAN[3] to classify posts into three different sentiment classes, Negative, Neutral and Positive. This model uses a variation of BERT[4] post-trained with reviews from Amazon and Yelp. Each post is transformed into an embedding sequence so as to apply sequence-pair classification to it. We define $x$ as the sequence embedding of a post:

$$x = [CLS]a_1, ..., a_m[SEP]t_1, ..., t_n[SEP] \quad (1)$$

where $a_1, ..., a_m$ are tokens of an aspect, $t_1, ..., t_n$ are tokens of words in a post, [*SEP*] is a separation token, and [*CLS*] is a token representing the classification value of the whole sequence embedding. Softmax is applied to the embedding in order to output probabilities for each sentiment class towards the aspect in post. This model has an accuracy of 74.3% when applied to a Reddit and 4chan political dataset [3] and 81.1% when evaluated on posts with less than 23 words.

**3. Aspects and keywords.** Our goal is to study the discourse surrounding GMOs. In order to capture the full discussion surrounding GMOs, we expand our initial keyword set to include strongly related topics. In our study, our keyword set includes:

- "CRISPR": The well-known method to edit genes.
- "Biofortifying": Selective breeding or genetic modification of crops.
- "Genomic": Study of genes and their influence on plant growth and development.
- "Bio-technology": The technology of modification of organisms.
- "Gene-editing": General term for the DNA modification of organisms.
- "Plant Breeding": A science-driven process for developing new plant varieties.

A post is considered relevant to GMOs if it contains any of the above keywords.

**4. Topic Extraction Tools.** We use the following three methods to extract meaningful information from our dataset.

**Global Topic Representation.** First each document is transformed into an embedding in vector space using the pre-trained embedding Sentence-Bert. UMAP is used to reduce the dimensionality of the embeddings, since it retains local and global features more effectively than PCA and t-SNE. HDBSCAN is then used to cluster the document embeddings together, accounting for any noise in the dataset which it does not put into any of the clusters. This method is highly effective for generating quality topic representations. Each cluster of documents is concatenated and considered as a single document. Then the following class-based TF-IDF calculation is applied:

$$W_{t,c} = tf_{t,c} * log(1 + \frac{A}{tf_t}) \quad (2)$$

Where class $c$ is a collection of concatenated documents in a cluster and $tf_{t,c}$ represents the frequency of term $t$ in cluster $c$. The inverse document frequency is calculated as the logarithm of the average number of terms in the document class $A$ divided by the frequency of term $t$ across all clusters of documents.

**Temporal Topic Modeling.** In order to take into account representations of topics appearing differently over time we also calculate the c-TF-IDF of topics at each timestep *i*. Global IDF values are calculated first across the entire corpus of documents, then term frequency at timestep *i* is multiplied against the global IDF.

$$W_{t,c,i} = tf_{t,c,i} * log(1 + \frac{A}{tf_t}) \quad (3)$$

This creates both global and temporal representation of each cluster of documents. The temporal representation is used to expand the keyword set of the global representation in our work. For example, one of our global topics is Climate Change and temporal aspects reveal 'fracking', 'greenhouse gasses', 'extractive farming', 'green revolution', and 'methane' as words users also use to represent the concept of Climate Change.

**Non-Negative Matrix Factorization.** Once we have our clusters of documents we apply NMF in order to get the

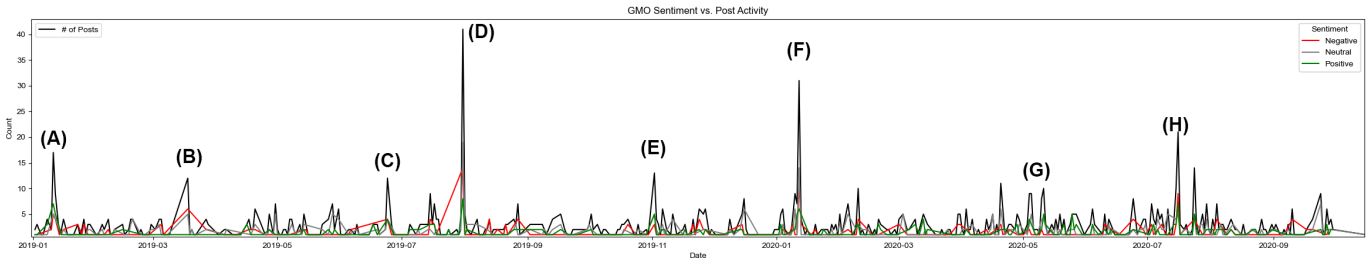


**Fig. 2:** Post Sentiment towards GMOs vs. Time, where spikes in post count align with real-world events. **Positive** spikes: (A) India court rules Monsanto can claim patents on GM cotton seeds. (C) USDA announces new rules requiring GMO disclosure labels on foods. (E) USDA finalizes rules letting companies determine if their biotech crops pose plant-pest risk. **Negative** spikes: (B) Genetically engineered salmon approved for consumption by FDA. (D) Roundup lawsuit against Monsanto announced. (F) USDA GMO labelling mandate takes effect. (G) FDA launches "Feed your Mind" Genetic Engineering education initiative. (H) Court upholds Monsanto verdict, slashes award by millions.

top coherent narratives fueling the concepts. Our document corpus is now more similar, the conditions necessary for this unsupervised algorithm to perform optimally.

We first create a term-matrix using a TF-IDF vectorizer to weigh unique terms higher. Our TF-IDF Vectorizer is modified to consider $n$-gram $\in [2, 4]$. Doing so will capture meaningful semantics of an argument or viewpoint. The NMF algorithm decomposes this matrix into two, one with topics or clusters of words discovered from the term matrix and the other with weights from each of the topics. The NMF continues to iterate in an unsupervised manner and solves for the objective function which is the Frobenius norm: $||A||^2_{Fro} = \sum_{i,j} A^2_{ij}$

## III. Case Study

We develop *Forumlyze*, an approach to systematically understand the prevailing concepts, narratives, and context as they exist in the discourse surrounding GMOs in online forums. Our approach consists of three major components: (a) algorithmically extracting the most popular concepts surrounding an issue, (b) qualifying and quantifying sentiment of these concepts, and (c) characterizing the meaningful relationship between an issue and concepts through context.

### A. Concept Extraction

First, we collect all the posts that contain at least one of our keywords of interest. This reveals that 0.8% of all the combined community posts have explicit mentions of the GMO issue. In order to capture the entire conversation we apply BERTopic over these posts, revealing class based TF-IDF importance scores in equation 2 for each of the revealed concept. We find that the concept of 'Climate Change' has a importance score of 0.009, 'Soil Science' 0.0102 and 'Monsanto' 0.041.

Next, we calculate c-TF-IDF of equation 3 for each timestep in our dataset. This expands our keyword set to include terms that may only exist at certain time periods but are still important to the concept. For example, our concept Monsanto has the following additional keywords: 'bayer', 'monopoly', 'roundup', 'corteva', 'cancer', 'class action', and 'lawsuit'.

With our new concept keywords, we revisit our dataset and search for additional posts that directly reply to a post mentioning GMOs, or relabel existing GMO posts to include the concept it's discussing as well. Furthermore, we consider posts that reply directly to GMO posts but that do not contain any of the new concept keywords as part of a concept of GMOs. This search expands our dataset by 941 posts. If a user is directly replying to a post that mentions GMOs, but their post itself only mentions a concept we assume that the context is still relevant and argue that it is important to take under analysis.

### B. Issue and Concept Characteristics

**1. Basic Statistics.** We first observe the following, GMOs, dominates the discussion with 847 unique users making 1474 posts about the issue. Monsanto is the next most popular with 309 posts from 204 users, then Soil Science with 190 posts from 160 unique users and finally Climate Change with 121 posts and 99 unique users. The total number of posts relating to GMOs and all its discovered concepts comes out to 2,094 or almost 1% of all posts made on the four communities.

**2. Community Popularity.** The concepts 'Monsanto', 'Climate Change', and 'Soil Science' are most popular within the Farming community with 77%, 53%, and 56% of all concept discussion happening within r/farming respectively. The issue of GMOs is most popular within the r/farming community, with 66% of posts originating from there. Additionally, 20% of posts originate from r/vegetablegardening.

### C. Sentiment Analysis

Using our Aspect-based Sentiment analysis tool, we analyzed the sentiment of all the collected posts about GMOs, as well as each subset of posts that mention related concepts.

**1. GMO Polarity.** GMOs as an issue is divided as being 31% Positive and 42% Neutral. The concept Monsanto is 39% Negative, and 40% Neutral, Climate Change is 30% Negative and 33% Neutral and Soil Science is 42% Positive and 38% Neutral. This nearly even distribution amongst the three sentiment classes supports evidence of polarized discussion surrounding GMOs, and when investigated further, GMO-related concepts maintained the observed polarity as well.

**2. Sentiment over Time.** By plotting a sentiment continuous probability density curve over time for each concept we can observe changes in sentiment as an evolutionary representation. The concept of Climate Change has a net decrease in Positive sentiment, beginning with a probability of 40% in January 2019 and decreasing to 30% by September of 2020. However between that time there are two peaks in June 2019 and June 2020. One explanation could be that June 2019 recorded the second hottest temperatures in the past 140 years and June 2020 was recorded as being the third hottest. While there were no drastic increases in Climate Change discussion during this time,

sentiment did change, indicating a shift in sentiment that was not an effect of post frequency.

For the concept Monsanto, Positive and Negative sentiment increased slightly as seen in Figure 1. However at the end of July 2019, citizens who successfully sued Monsanto over allegations that their product Roundup caused cancer, had their award decreased by nearly $50 Million USD. A gradual decrease in Positive sentiment density and subsequent gradual increase in Negative density was observed over the next few months. These density values ended up back at their original values a few months later, but a rapid increase was also observed at the end of July in 2020, an explanation for this, and other spikes in sentiment can be seen in Figure 2. Initial manual inspection of the content of these posts seems to suggest there is a relationship, but establishing more than an empirical correlation will require additional tests.

**D. The Context of Concepts.** Now that we have identified our concepts, their relationship to GMOs and the sentiments of these concepts we briefly investigate their respective context. We define the context broadly as the framing of the main representation of a concept. We use NMF since our concepts have already narrowed down our dataset and separated the posts into class-based clusters. We identify the top (2,4)-gram narratives for each of our concepts.

| **Monsanto** |
|---|
| pesticides industry shoves, consumer news, food antibiotics |
| honestly believe monsanto companies, theyve bough control |
| main ingredient, partly responsible total |
| skin cancer assume monsanto, lawsuit lawyer |
| feed kids healthy, cancer monsanto humans, proof causes |
| evil monsanto, farming price, bayer organic settlement |
| roundup ready, okay sponsored, monsanto propaganda |
| **Climate Change** |
| biodegradable plastics, rarely recyclable, small pieces pollute |
| climate change, importance explaining science question |
| small homesteaders, environmental policies |
| pesticides pollution, true hybrid wheat |
| believe science, climate change, genetic engineering |
| protesters gmo, fracking climate change denial |
| environmental conditions protection, chemicals glyphosate |
| **Soil Science** |
| tech basic, fundamental covered |
| soil repot, veggies planted, amended bone blood |
| masters degree, interested know worked |
| soil science degree, potting days |
| gmos soil science, organic fertilizers |
| chat agronomy grower, state position, cannabis legal |
| raised beds weather, france gmos, allowed europeans country |

## IV. Related Work

This study differs from previous efforts because (a) it focuses agricultural forums, (b) goes beyond word "counting" by identifying related concepts to an issue of interest, and (c) by going deeper into understanding the context of these related concepts. In addition, we use a comprehensive set of tools including, sentiment analysis leveraging BERT and BERTopic.

To the best of our knowledge, no prior work has focused on mining agricultural forums. By contrast, most studies have analyzed Twitter as we discuss below. We argue that Twitter is more of an announcement and "posturing" forum, while a discussion forum captures in-depth opinions of its users. We keep this discussion to a minimum due to space limitations.

**Forum analysis:** Recent efforts [5], [6] study political discourse sentiment, and the posts that drive discussion on Reddit.

**GMO analysis & Topic Modeling:** There are several studies focusing on Twitter datasets analyzing public discussion of GMO. These works [7], [8] investigate popular topics and conversations. [9] analyzes the different geographic sentiment of GMO themes. [10] uses LDA topic modelling to extract thematic discussions, and apply sentiment analysis to discover the differences associations and perceptions.

## V. Conclusion

The key contribution of our work is *Forumlyze*, a systematic framework to holistically understand the GMO debate in online forums. Our approach consists of three components: (a) collecting and systematizing relevant posts to the issue (b) discovering concepts surrounding the issue using temporal class-based topic modelling and (c) identifying the main narratives behind the concepts using Non-negative Matrix Factorization.

We argue that our initial results are promising. Our follow up work will provide an extensive and detailed "map" of: (a) topics, (b) user sentiment, and (c) temporal considerations and evolution of this debate.

## VI. Acknowledgments

This work was supported by CDFA grant 2021 Specialty Crop Block Grant Program H.R. 133 No. 000705282.

## References

[1] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.
[2] Pushshift. https://pushshift.io/.
[3] J. Tachaiya, J. Gharibshah, K. E. Esterling, and M. Faloutsos, "RAFFMAN: measuring and analyzing sentiment in online political forum discussions with an application to the trump impeachment," pp. 703–713, 2021.
[4] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification," *arXiv preprint arXiv:1908.11860*, 2019.
[5] J. Tachaiya, A. Irani, K. M. Esterling, and M. Faloutsos, "Sentistance: quantifying the intertwined changes of sentiment and stance in response to an event in online forums," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 361–368.
[6] B. D. Horne, S. Adali, and S. Sikdar, "Identifying the social signals that drive online discussions: A case study of reddit communities," in *2017 26th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2017, pp. 1–9.
[7] C. D. Wirz, E. L. Howell, D. Brossard, M. A. Xenos, and D. A. Scheufele, "The state of gmos on social media: An analysis of state-level variables and discourse on twitter in the united states," *Politics and the Life Sciences*, vol. 40, no. 1, pp. 40–55, 2021.
[8] J. Ji, M. Robbins, J. D. Featherstone, C. Calabrese, and G. A. Barnett, "Comparison of public discussions of gene editing on social media between the united states and china," *Plos one*, vol. 17, no. 5, p. e0267406, 2022.
[9] K. Munro, C. M. Hartt, and G. Pohlkamp, "Social media discourse and genetically modified organisms," *The Journal of Social Media in Society*, vol. 4, no. 1, 2015.
[10] I. Jun, Y. Zhao, X. He, R. Gollakner, C. Court, O. Munoz, J. Bian, I. Capua, and M. Prosperi, "Understanding perceptions and attitudes toward genetically modified organisms on twitter," in *International Conference on Social Media and Society*, 2020, pp. 291–298.